Thinking Skills and Creativity 48 (2023) 101255

Contents lists available at ScienceDirect

Thinking Skills and Creativity

journal homepage: www.elsevier.com/locate/tsc

The development and validation of the scale of design thinking for teaching (SDTT)

Yuyang Cai [a], Yan Yang [b,*]

[a] *Centre for Language Education and Assessment Research (CLEAR) & School of Languages, Shanghai University of International Business and Economics, 1900 Wenxiang Road, Songjiang District, Shanghai, China*
[b] *School of Languages, Shanghai University of International Business and Economics, 1900 Wenxiang Road, Songjiang District, Shanghai, China*

ARTICLE INFO

Keywords:
Convergent validity
Design thinking
Discriminant validity
Instruction quality
Measurement invariance

ABSTRACT

Design thinking has exhibited significant strength in instruction optimization. However, relevant research has been limited due to the unavailability of measurement for teachers' design thinking. The current study developed and validated the Scale of Design Thinking for Teaching (SDTT). Drawing on responses from 1,018 K-12 teachers in Mathematics and English as a foreign language, we cross-validated the scale's measurement quality and then examined its predictive validity for instruction quality. Results suggested the SDTT had four factors (*problematizing, ideating, prototyping,* and *testing*) with acceptable reliability, convergent validity, discriminant validity, strong measurement invariance across subject and gender groups, and good predictive validity for instruction quality. The SDTT can be utilized as a list of concrete and operable strategies for teachers to apply in instructional practice and as a diagnostic tool for teacher educators in teacher training programs. Limitations and suggestions for future research were discussed.

1. Introduction

Student learning largely hinges on the effectiveness of teachers' instruction (Motallebzadeh et al., 2018). Design thinking, a human-centered problem-solving approach that emerged from the designer profession, has been recognized as an essential process in designing these instruction activities (Henriksene et al., 2018; Koh et al., 2015; Madson, 2021). Design thinking empowers teachers to enhance instructional effectiveness by taking students' perspectives, identifying the problem, and evaluating and refining the teaching plan (Henriksen et al., 2018). Empirical studies have testified to the beneficial impact of design thinking on teachers' instruction and student learning (Atchia, 2021; Goldman & Kabayadondo, 2017; Wu et al., 2019).

Various attempts have been made in the education sector to help teachers develop their design thinking (Crites & Rye, 2020; d. school, 2018). However, the lack of instruments to measure teachers' design thinking has limited research and practice in design thinking for instruction. To address this gap, the current study aimed to develop and validate an instrument to measure teachers' design thinking. The newly developed measures would facilitate large-scale empirical research that examines the relation between design thinking and instructional practice and how this relation can be subject to other variables. Furthermore, this self-report design thinking measurement is expected to help pre-and in-service teachers to understand the ingredients of quality instruction. It will also likely provide helpful information for supervisors and teacher educators when developing their teacher education curricula.

* Corresponding author.
E-mail address: yangyan_alexyang@163.com (Y. Yang).

https://doi.org/10.1016/j.tsc.2023.101255
Received 2 August 2022; Received in revised form 26 January 2023; Accepted 26 February 2023
Available online 28 February 2023
1871-1871/© 2023 Elsevier Ltd. All rights reserved.

2. Literature review

2.1. Design thinking

Design thinking has been regarded as a solution-based skill or approach used by experienced designers to creatively address complex problems (Dorst, 2011; Razzouk and Shute, 2012). This approach emphasizes that problem-solving should be anchored in understanding and fulfilling human needs (Brown, 2008). Drawing on this understanding, the problem is identified, and a wide range of ideas is generated. This is followed by a field test of the preliminary solution built on alternative ideas, possibly leading to several iterations of the procedures to determine the optimal solution (d.school, 2018).

Design thinking has gained increasing popularity in diverse professions, including business management (Sarooghi et al., 2019), healthcare (McLaughlin et al., 2019), and social innovation (Goi & Tan, 2021), to list but a few. Importantly, design also lies at the core of education, where educators strive for students' better learning outcomes through pedagogy design (Cross, 2001; Henriksen et al., 2018). The problems educators confront are similar to those that designers aim to address, that is, wicked problems that are ill-defined without one straightforward solution (Buchanan, 1992). The difficulties in teaching practice are complicated due to various issues, such as pedagogical goals, students' cognitive capacity, learning motivation and engagement, teacher-student relationship, and so forth. Given this analogy, design thinking has been widely heralded as an essential problem-solving approach for teachers to aid pedagogy design and achieve quality teaching (Koh et al., 2015).

Different design thinking models have been proposed to cater to specific professional fields (Lin et al., 2020). In teacher education, the Stanford Design Thinking Model has been widely adopted, which specifies five dimensions, i.e., *emphasize, define, ideate, prototype*, and *test* (d.school, 2018). While design thinking has been extensively included in teacher training and received considerable attention in educational research (Henriksen et al., 2018), large-scale empirical studies are limited due to a lack of valid scale measuring design thinking in the teaching context.

In educational studies, however, scales have been developed to measure the design thinking traits of students, which we find informative for developing scales to measure teachers' design thinking. For example, Blizzard et al. (2015) developed an 18-item questionnaire measuring design thinking traits and validated it among 6772 college students studying engineering in the United States. Results of EFA (exploratory factor analysis) suggested five factors underlying nine items: collaboration, experimentalism, optimism, feedback-seeking, and integrative thinking. Results of linear regression suggested the validity of these factors in predicting other constructs, including learning achievement. In the discipline of business, Roth et al (2020) validated a 20-item questionnaire for design thinking among 160 business engineering students. Results of MIMIC (multiple-indicator, multiple-cause model) suggested five dimensions: user as an information source, user as co-developer, problem framing, prototyping, and iteration. A common feature of these scale development studies is their exploratory nature. For this reason and many others (e.g., different purposes and contexts), different studies have identified different traits. Although these identified traits appear inclusive, they are atheoretical and risk over-representing the design thinking construct.

More recently, Tsai & Wang (2020) built their work on the Stanford model and developed a scale of 20 candidate items for design thinking for studying computer programming. Results of EFA with 350 Taiwanese adolescents retained 18 items and confirmed four factors: Ideate, Prototype, Empathize, and Define. However, since it is hardly possible for computer programming learning in secondary education to involve *test* dimension as explained by the authors, this dimension was omitted in the development process, which somehow risks the under-representation issue.

In short, endeavors in the measurement development of design thinking have evolved from the atheoretical approach to being more theory-based. While the Stanford design thinking model seems to provide a useful conceptual framework, to our knowledge, it has rarely been used in scale development. Therefore, our study adopted the Stanford design thinking model as the theoretical basis, which will be elaborated in the following section.

2.2. Factors of design thinking

According to the five-mode Stanford Design Thinking Model, when teachers apply design thinking, they *empathize* with students, d*efine* the problem, *ideate* the solutions, *prototype* the solutions or schemes, and *test* the efficiency of these strategies (d.school, 2018). For the sake of convenient communication, we convert the verbs of the five modes into their gerunds as *empathizing, defining, ideating, prototyping*, and *testing.*

Empathizing. Through empathizing, teachers gather comprehensive information about what students need and want and how they feel during learning (d.school, 2018; IDEO, 2016; Radford University, 2013). This step aligns with needs analysis in task-based language teaching (TBLT; Ellis, 2017; Long, 2014). According to TBLT, tasks, content, and strategies in the classroom should be designed based on learners' needs, and this can be achieved through interactions with specific learners (Long, 2014). In classroom teaching, teachers should try to follow the thread of students' cognitive process, a teacher quality known as epistemic empathy (Jaber et al., 2018). Besides, teachers should also notice the contextual issues involved in the learning process related to motivational or affective variables (Robertson et al., 2015). To empathize with students, various strategies can be applied by the teacher, such as interviewing students for direct responses, observing learners' behaviors for implicit information, and taking learners' perspectives by imagination (English, 2016).

Defining. In this mode, teachers explicitly state the most meaningful and actionable problem with the student-related information obtained from empathizing. Data resulting from empathizing will be studied and integrated to rule out irrelevant possibilities and pinpoint the critical problem. This mode can also find its echoes in other research paradigms related to teaching, namely, the identification of tasks in TBLT, or the selection of specific problems in problem-based learning or inquiry-based learning, in that all of them require teachers to locate the pain points and search for the crux of the problematic situation (Hmelo-Silver, 2004; Long, 2014). When defining, teachers should hold back their preconceptions to avoid stereotypes likely to undermine the objectivity and authenticity of the problem (d.school, 2018; Liedtka, 2015; Long, 2014). When defining the problem, teachers should also consider the constraints in classroom teaching, such as top-down teaching guidelines, length of courses, and size of the class (Crites & Rye, 2020).

Ideating. Ideating refers to the process in which teachers generate a wide range of possible ideas or ways that might contribute to problem-solving (Crites & Rye, 2020; d.school, 2018). This mode encourages creativity and open-mindedness in teachers and requires them to utilize their divergent thinking to help themselves think wide (d.school, 2018). All teaching methods and strategies teachers are acquainted with can be recorded to compile the initial pool of ideas. Instead of individual efforts, teachers can utilize multiple ways to generate ideas by collaborating with colleagues or observing good classroom teaching (Wu et al., 2019).

Prototyping. During prototyping, teachers quickly shape the ideas into a preliminary teaching plan after narrowing down the ideas created in the ideating mode. Prototyping is not a perfect final solution but good enough to be put into trials in the Testing mode (d.school, 2018). Teachers must use convergent thinking and consider variables in authentic learning situations to create the teaching plan. The teaching plan would be a ready-to-test model instead of a perfect one (Crites & Rye, 2020; Nation & Macalister, 2019). Teachers may build the model in an outline, a mind map, or tangible objects, which allow them to interact with, deeply understand, anticipate the potential drawbacks, and revise it immediately if necessary (d.school, 2018).

Testing. Testing is teachers' chance to 'gather feedback, refine solutions, and continue to learn' (d.school 2018, p. v) about their students. In this mode, teachers conduct a pilot study to test the prototype with target learners in the actual learning environment. Based on the solution's evaluation, teachers decide whether to adopt or redesign it. In the testing mode, teachers evaluate their teaching plans with triangulated resources, such as students' implicit and explicit feedback, performance, and so forth. This evaluation process demands teachers' metacognitive skills to move beyond the role of insiders to reflect on their teaching practice. This reflection echoes the notion of "reflective teaching" in education literature (Zeichner & Liston, 2013). To harvest maximum benefit from field tests, teachers must foster the willingness to accept failures and learn from trials and errors (Razzouk & Shute, 2012).

2.3. *Design thinking and teaching*

Design thinking has been touted as a structured guideline to help teachers integrate pedagogical knowledge and contextual issues, design practical and creative teaching activities and improve their confidence in teaching practice (Henriksen et al., 2018; Khodabakhshzadeh et al., 2018). Empirical research has demonstrated the benefits of design thinking to teaching practice in different domains, such as teacher education (Henriksen et al., 2018), science, technology, engineering, and mathematics (STEM) education (Atchia, 2021; Wu et al., 2019), and language education (Crites & Rye, 2020; Matsui, 2020).

Through a case study, Atchia (2021) observed that the teacher's use of design thinking efficiently enhanced students' engagement and development in their biology learning. The researcher found that the teacher revised her teaching and assessment strategies and adjusted resources under the design thinking guideline by considering students' feedback towards teaching. These adjustments eventually led to higher student engagement (Atchia, 2021).

Henriksen et al. (2018) documented a graduate-level teacher education course based on the Stanford model. The results demonstrated that both in-service and pre-service teachers acknowledged that design thinking practices equipped them with strategies to solve context-specific problems creatively. These strategies included appreciating empathy, tolerating uncertainty, and seeing teachers as designers (Henriksen et al., 2018).

Tseng et al (2019) examined the role of design thinking in fostering six pre-service teachers' Technological Pedagogical Content Knowledge during their web-conference teaching. Results of quantitative content analysis and qualitative analysis of interviews showed the conspicuous deployment of Pedagogical Content Knowledge by the pre-service teachers, suggesting the potential of design thinking in facilitating pre-service teachers' competence in solving contextual problems during online instruction, such as technical problems and students' attention problems.

In a case study, Crites & Rye (2020) explored the implementation of design thinking in the curriculum design for an English as a Foreign Language (EFL) program at a Columbia university. The researchers found that integrating design thinking into the curriculum design process improved teachers' design practices by making the design process more collaborative, creative, and efficient, and these advantages carried on to their course instruction.

The swath of studies in design thinking has shown the potential of design thinking for empowering teaching. However, most studies have been constrained to case studies or small-scale qualitative inquiries. A consequence of this constraint is that the results are usually idiosyncratic whose generalizability to a larger scale remains an open issue. To advance research and practice studies in teachers' design thinking, the current study aimed to develop and validate a scale measuring designing thinking for teaching. The present study attempted to develop the Scale of Design Thinking for Teaching (SDTT), evaluate its measurement quality, and assess the predictive validity of the SDTT for instruction quality.

3. Methods

3.1. *Scale construction*

Scale construction involved two steps. First, the authors reviewed the literature pertinent to design thinking extensively. Items were then generated by referring to existing design thinking questionnaires, which resulted in a pool of more than 30 candidate items.

Second, the researchers divided these items into five dimensions based on the Stanford model and translated them into Chinese with the help of a professor of EFL education and translation using the back-forward approach (Grisay, 2003, 2006). Afterward, a panel of seven members was formed, which comprised one professor of educational assessment and higher-order thinking (the first author), one professor in EFL education and translation, three lectures in foreign languages education (two in English and one in French), and two graduate students studying higher-order thinking in language education assessment (including the second author). The panel reviewed the items regarding their wording, clarity, content coverage, relevance to teaching practice, and order that might constrain teacher participants' responses. During this course, the survey items were online for the panel members to try out. After more than ten rounds of trying out and revising, the panel agreed on a pool of 29 items.

3.2. Instruments

SDTT. This 29-item scale contained five subscales: empathizing, defining, ideating, prototyping, and testing. The questionnaire asked teacher participants to rate the frequency of the behavior depicted by each item on a six-point Likert scale (1= "Never," 2= "Very Rarely," 3= "Rarely," 4= "Occasionally," 5= "Frequently," 6= "Always"). Since neutral responses might be interpreted as neutral, undecided, or avoiding responding, the neutral point was not adopted as it would bear the risk of ambiguity in data analysis (Krosnick et al., 2002). The instruction for responding was: "Please read each of the statements below carefully and recall your teaching practice, and then indicate the extent to which each statement reflects your teaching practice." A sample item is "(I try to) encourage students to talk about how they feel during conversations" (tapping into *emphasizing*). Please refer to Appendix A for details of the scale.

IQS (Instruction Quality Scale). The IQS was adapted from the OECD Teaching and Learning International Survey (TALIS) (Ainley & Carstens, 2018; OECD, 2018). TALIS defined instruction quality as a multidimensional construct that comprises five dimensions: classroom management, clarity of instruction, cognitive activation, feedback to students, and assessment strategies. We used a shortened TALIS teaching quality scale to reduce response workload by choosing one item from each of the five dimensions. The use of shortened questionnaires to reduce respondents' workload and enhance data response quality is not rare to find (Bento et al., 2019; Kim et al., 2018; OECD, 2019). For instance, OECD (2019) even used a single item to measure fluid intelligence, and the measure has been proved to be quite effective in capturing the primary feature of fluid intelligence (Bernardo et al., 2021).

In our study, the 5-item IQS asked teacher participants to evaluate their teaching practice and rate on a 6-point Likert scale (from 1 = "Strongly Disagree" to 6 = "Strongly Agree"). Item 1 measured classroom management ("I calm students who are disruptive," M = 5.52, SD=0.71), Item 2 addressed instruction clarity ("I set a goal at the beginning of instruction," M = 5.52, SD=0.66), and Item 3 was about cognitive activation (" I give tasks that require students to think critically," M = 5.03, SD=0.91), Item 4 tapped into feedback ("I provide written feedback on student work in addition to a mark " M = 5.27, SD=0.78), and Item 5 dealt with assessment ("I use a variety of assessment strategies," M = 5.20, SD=0.84). The internal consistency of the IQS was α=0.81, indicating good reliability. The measurement quality of the scale was also corroborated through CFA (see later text).

3.3. Data collection

The researchers collected data from 127 schools in a city in South China. Before data collection, consent was obtained from the education authority of the town, local schools, and participating teachers. The study recruited 1018 in-service teachers teaching EFL or Mathematics at the compulsory education level. Compulsory education in China includes primary (Grades 1–6) and junior middle (Grades 7–9) school levels that every child should complete as stipulated by China's Compulsory Education Law (National People's Congress, 2006). Among the participants, 65.2% were teaching at the primary school level (429 mathematics teachers and 235 EFL teachers), and 34.8% were teaching at the junior middle school level (181 mathematics teachers and 173 EFL teachers).

About 79.3% of the teachers were female, and their ages ranged from 20 to 59 (M = 34.80, SD = 9.39). Regarding their educational background, 76.9% were with a bachelor's degree, 9.5% with a master's degree, 13.5% with an associate degree, and 0.1% with a doctoral degree. Their years of teaching varied from 1 to 40 years (M = 11.56, SD = 10.11).

The SDTT and IQS used for data collection were delivered via the online platform Wenjuanxing (https://www.wjx.cn). The questionnaire included a cover letter informing the participants of the purpose of the study, volunteering participation, contents of the questionnaire, participating confidentiality ensuring, and the time needed to finish the questionnaire. Teachers who accomplished this survey were therefore recognized as giving consent to data collection.

3.4. Data analysis

Primary data analysis involved three steps: 1) cross-validation, 2) assessment of reliability assessment, convergent validity, discriminant validity, and measurement invariance, and 3) predictive validity.

All main analyses were run on Mplus version 8.5 (Muthén & Muthén, 1998–2020). When conducting EFA, CFA, and MIMIC analyses (Kline, 2015), we used the estimator of maximum likelihood with robust standard errors (MLR) due to its robustness to non-normal data (Satorra & Bentler, 2001) with the Geomin rotation method (Muthén & Muthén, 1998–2017). Model-data fit was assessed using multiple criteria: comparative fit index (CFI) and Tucker-Lewis index (TLI), root mean square error of approximation (RMSEA), and standardized root mean square residual (SRMR). Models with CFI and TLI values of 0.95 or above and with RMSEA and SRMR values of 0.05 or less are considered to have a good fit, whereas models with CFI and TLI values above 0.90 and RMSEA and SRMR values smaller than 0.08 are considered to have an acceptable fit (Mueller et al., 2019).

3.4.1. Step1: cross-validation

To conduct cross-validation, we first randomly split the data into two halves. Afterward, we conducted EFA with Subsample 1 and confirmatory factor analysis (CFA) with Subsample 2. The EFA was conducted to examine the dimensionality, and factorial structure of the SDTT, and the CFA was to evaluate the stability of the factorial structure established through EFA with a different dataset. Factors with an eigenvalue above one (Brown, 2015) and model fit indices reported by M*plus* were consulted when deciding the number of factors to be extracted (Aichholzer, 2014). It is recommended that factor retention decisions be made based on multiple criteria (Hayton et al., 2016). Therefore, we also conducted parallel analysis (PA), very simple structure (VSS) criterion, and Velicer's minimum average partial (MAP) to assist factor retention decision.

3.4.2. Step 2: assessment of reliability, convergent validity, discriminant validity, and measurement invariance

Reliability refers to the internal consistency of the items in measuring the construct underlying a scale. The reliability level for each subscale was evaluated by Cronbach's alpha (α), McDonald's omega (ω), and Coefficient H. Values of Cronbach's alpha and McDonald's omega above 0.70 indicate acceptable reliability (Cortina, 1993; Dunn et al., 2014), while Coefficient H was calculated to show the maximal reliability (Hancock & Mueller, 2001; McNeish, 2018).

Convergent validity refers to the extent to which the items within a scale measure the same construct (Cunningham et al., 2001). The evaluation of convergent validity was based on two statistics: average variance explained (AVE) and composite reliability (CR). Good convergent validity is determined if AVE is above 0.5 or a CR value is above 0.7 (Fornell & Larcker, 1981; Hair et al., 2019).

Discriminant validity demonstrates the extent to which a construct distinguishes itself from others. We adopted a fourfold approach to examine discriminant validity (Walker et al., 2015). First, the AVE value of each factor was compared with the squared correlations (r^2) between the factor and other factors. An AVE value larger than the squared correlations would suggest adequate discriminant validity between the factors (Fornell & Larcker, 1981). Factors that failed to satisfy this criterion were considered problematic for further inspection.

Second, a competing model with the correlation of problematic factors fixed to one was constructed and compared with the original correlated-factor model using the chi-square difference test. A significant difference between the two models would suggest sufficient discriminant validity (Kenny, 2016). For model comparison, as the Satorra–Bentler chi-square provided by M*plus* with the MLR estimator is not suggested to be directly used for the chi-square difference test (Satorra & Bentler, 2010), the chi-square difference test was computed using the Satorra–Bentler chi-square, scaling correction factor, and degree of freedom as suggested by Satorra & Bentler (2010).

Third, another competing model with problematic factors combined into one factor was compared with the original correlated-factor model. As the two models were non-nested, the comparison was made based on the model fit indices of each model instead of comparing their chi-square difference (Kenny, 2016). A better fit of the original correlated-factor model would suggest sufficient discriminant validity.

Fourth, the discriminant validity of the scale was also examined by inspecting the confidence interval around the correlation between the problematic factors (correlation estimates and standard error). If the range does not cover the value of one, the two factors are considered distinct, and sufficient discriminant validity is determined (Anderson & Gerbing, 1988).

Furthermore, we also tested whether the factorial structure holds across different subgroups of the same population (Van de Schoot et al., 2012). In our case, we conducted two-group CFA using the whole sample to evaluate the measurement invariance of the SDTT across subject and gender groups. To deal with the significant gender distribution imbalance (females = 79.3%, quite a normal gender distribution of K-12 teachers in China), we followed the method suggested by Yoon & Lai (2018). When assessing MI, we tested the following models in a hierarchical order: a configural model (unconstrained model), a weak invariance model (factor loadings constrained equal), and a strong invariance model (item intercepts constrained equal) (Widaman & Reise, 1997). An invariance level was determined if the more complex model had a change in CFI smaller than 0.01 (Cheung & Rensvold, 2002) and a change in RMSEA smaller than 0.015 (Chen, 2007).

3.4.3. Step 3: predictive validity

Drawing on the resultant SDTT cross-validated by the steps above, we examined the predictive validity of teachers' design thinking for instruction quality using the whole sample. First, a bivariate correlation analysis was conducted among design thinking subscales (represented by their means) and instruction quality to provide the basis for the following analyses. Second, CFA was performed with the whole sample to ensure the measurement quality of the shortened IQS. Third, MIMIC was conducted to examine the predictive validity of design thinking for instruction quality.

4. Results

4.1. Demographic information

Table 1 shows the demographic information of teachers in the two split samples. As shown, the two randomly split subsamples had similar means in age and years of teaching experience. The two random subsamples also had similar ratios in teachers of Mathematics and English. Regarding teaching levels, Subsample 2 had a slightly larger percentage of primary school teachers (68.0%) than Subsample 1 (62.5%). Regarding teacher education backgrounds, Subsample 2 had a slightly larger ratio of teachers with a Bachelor's degree (78.6%). To summarize, the distributions of demographic factors across the two subsamples were similar in most variables.

4.2. Cross-validation

The cross-validation was done in two steps: EFA with the first subsample and CFA with the second subsample. When conducting EFA on Mplus, a researcher must set the minimum and the maximum number of factors to be extracted. In line with the literature and our original design, we set the minimum number of factors as one and the maximum number as five. Table 2 shows the model fit indices for the first round EFA models containing one to five factors, respectively. As shown, the model fit improved as the number of extracted factors increased, and the fit indices reached an acceptable level until the five factors were extracted. Moreover, four of the five factors had eigenvalues larger than 1: 14.30, 1.82, 1.35, 1.15, and 0.93, respectively. In addition, the results of PA, VSS, and MAP suggested retaining two factors, one factor, and four factors, respectively. The one-factor structure could be understood as all items reflecting the same construct, while the two-factor structure had severe cross-loadings that were difficult to interpret theoretically. Furthermore, according to the model fit indices from EFA, both one-factor and two-factor structure were far from the acceptable level. Considering quantitative criteria, interpretability, and theoretical expectations, we finally decided to retain four factors.

We then inspected inappropriate items with weak factor loadings (items with loadings smaller than 0.30 on any factor) and strong cross-loadings (similar loadings) on at least two factors and conceptual relevance between factors and their loaded items. We excluded one inappropriate item at a time until each item was loaded on the relevant factor and no severe cross-loadings were displayed.

In the end, 17 items were retained, and a four-factor structure was determined: *problematizing* that combined *empathizing* and *defining* (6 items), *ideating* (4 items), *prototyping* (3 items), and *testing* (4 items). The eigenvalues for the four factors were 8.77, 1.14, 1.06, and 1.01, respectively, all above the rule-of-thumb value of 1. Besides, this model reached a good fit: CFI = 0.981, TLI = 0.965, RMSEA (90% CI) = 0.041 (0.030, 0.052), and SRMR = 0.022. The factor loadings of the final EFA solution are shown in Table 3.

The four-factor structure obtained from EFA was further cross-validated using Subsample 2. The model fit the data well: CFI = 0.961, TLI = 0.954, RMSEA (90% CI) = 0.048 (0.040, 0.056), and SRMR = 0.042. Fig. 1 presents the standardized estimates of the structural model. The correlations between the factors ranged from $r = 0.68$ (between *problematizing* and *prototyping*) to $r = 0.79$ (between *problematizing* and *testing*). Within each factor, the loadings were all 0.56 and above, indicating sufficient factor loadings of items with each factor.

4.3. Reliability, convergent validity, discriminant validity, and measurement invariance

Table 4 shows the statistics for the reliability, convergent validity, and discriminant validity calculated based on the CFA results with Subsample 2. As shown, Cronbach's alphas for each factor (from 0.86 to 0.88) and omega (from 0.82 to 0.88) are all larger than the cut-off value of 0.70, while Coefficient H ranged from 0.86 to 0.94, suggesting good reliability for each subscale.

Convergent validity was assessed based on the CR and AVE values. The CR values ranged from 0.84 to 0.88, all above the recommended threshold of 0.70 for good convergent validity (Hair et al., 2019). The AVE values of four factors were all greater than 0.50 (from 0.66 to 0.77) except for *problematizing* (AVE = 0.47), which roughly met the criteria of good convergent validity. As AVE is a relatively conservative criterion, it has been recommended that CR alone could be used for deciding convergent validity when an AVE falls below 0.50 (Fornell & Larcker, 1981). To summarize, the SDTT showed adequate convergent validity.

The assessment of discriminant validity consisted of four types of analyses. First, the AVE value of a factor was compared with the squared correlation between the factor and another corresponding factor. As shown in the right block of Table 4, the squared correlations between two pairs of factors exceeded the AVE value of the related factor (AVE = 0.47): the squared correlation between *problematizing* and *ideating* ($r = 0.57$) and that between *problematizing* and *testing* ($r = 0.62$). The results indicated the two pairs of factors (i.e., *problematizing* and *ideating, problematizing* and *testing*) might need further examination to determine the discriminant validity of the scale.

Second, the original four-correlated-factor model was compared with two competing models, and the fit statistics are shown in the discriminant validity block in Table 2. The first competing model forced the correlation between *problematizing* and *ideating* to one (Model 1.1), and the second competing model fixed the correlation between *problematizing* and *testing* to one (Model 1.2). The fit

Table 1
Demographic information.

Variable	Subsample 1	Subsample 2
Gender	77.6% female	80.9% female
Age	$M = 35.07$ ($SD = 9.67$)	$M = 34.53$ ($SD = 9.10$)
Years of teaching experience	$M = 11.72$ ($SD = 10.35$)	$M = 11.40$ ($SD = 9.88$)
Subjects		
English	201 (39.5%)	302 (40.7%)
Mathematics	308 (60.5%)	207 (59.3%)
Teaching level		
Primary school	318 (62.5%)	346 (68.0%)
Junior middle school	191 (37.5%)	163 (32.0%)
Education background		
Associate degree	74 (14.5%)	63 (12.4%)
Bachelor's degree	383 (75.2%)	400 (78.6%)
Master's degree and above	52 (10.2%)	46 (9.0%)

Notes. M = mean, *SD* = standard deviation.

Table 2
Model fit indices.

Model	S-Bχ^2	*df*	RMSEA (90% CI)	SRMR	TLI	CFI
First-round EFA (n_1= 509)						
1-factor	1890.375*	377	.089 (0.085, 0.093)	.064	.764	.781
2-factor	1527.725*	349	.081 (0.077, 0.086)	.049	.801	.829
3-factor	1190.689*	322	.073 (0.068, 0.077)	.041	.841	.874
4-factor	834.387*	296	.060 (0.055, 0.065)	.033	.893	.922
5-factor	486.037*	271	.039 (0.034, 0.045)	.022	.953	.969
Final round EFA (n_1= 509)						
1-factor	795.190*	119	.106 (0.099, 0.113)	.062	.771	.800
2-factor	556.781*	103	.093 (0.086, 0.101)	.052	.823	.866
3-factor	311.013*	88	.071 (0.062, 0.079)	.040	.898	.934
4-factor	137.434*	74	.041 (0.030, 0.052)	.022	.965	.981
CFA (n_2= 509)						
4-factor	247.049*	113	.048 (0.040, 0.056)	.042	.954	.961
Discriminant validity (n_2= 509)						
Model 1.1	416.874*	114	.072 (0.065, 0.080)	.047	.896	.913
Model 1.2	406.351*	114	.071 (0.064, 0.079)	.051	.900	.916
Model 1.3	420.582*	116	.072 (0.065, 0.079)	.047	.897	.912
Model 1.4	411.093*	116	.071 (0.063, 0.078)	.052	.901	.915

* $p < .001$; S-B χ^2 = Satorra–Bentler chi-square; *df* = degrees of freedom; RMSEA=root mean square error of approximation; SRMR=standardized root mean square residual; TLI=Tucker–Lewis index; CFI=comparative fit index.

Table 3
Factor loadings (n_1 = 509).

Item	F1	F2	F3	F4
02 try to know about the students' learning needs.	.410			
03 encourage students to talk about how they feel during conversations.	.547			
05 observe students to understand them better.	.567			
06 try to understand students in multiple ways.	.582			
08 specify the difficulties in classroom instruction.	.655			
10 constantly summarize the characteristics of quality teaching.	.599			
14 generate new teaching ideas by brainstorming.		.316		
15 explore new ideas by reflecting on my teaching practice.		.679		
16 learn about teaching approaches from experienced colleagues.		.947		
17 get new teaching ideas by observing other teachers' classes.		.799		
20 go over a teaching idea with examples.			.492	
21 generate flow charts or outlines to visualize the instruction procedure.			.946	
22 draft a teaching plan to better understand a teaching approach.			.820	
25 test the teaching plan in an actual classroom setting.				.752
26 pay attention to students' responses when I test a teaching plan.				.576
27 modify a teaching plan according to students' feedback.				.876
28 reflect on teaching practice to improve the teaching plan.				.627

Note. Factors 1–4 = problematizing, ideating, prototyping, and testing; Factor loadings lower than 0.30 were not displayed; All loadings are significant at $p < .05$.

statistics for Model 1.1 were CFI = 0.913, TLI = 0.896, RMSEA (90% CI) = 0.072 (0.065, 0.080), SRMR = 0.047, and those for Model 1.2 were CFI = 0.916, TLI = 0.900, RMSEA (90% CI) = 0.071 (0.064, 0.079), SRMR = 0.051, both worse than the fit statistics of the original four-correlated-factor model: CFI = 0.961, TLI = 0.954, RMSEA (90% CI) = 0.048 (0.040, 0.054), SRMR = 0.042.

We made a further comparison between the original model, Model 1.1, and Model 1.2. The chi-square test showed that the two competing models differed significantly from the original model ($p < .001$). These results supported discriminant validity between *problematizing* and *ideating* and between *problematizing* and *testing*.

Third, we compared the original model with two other competing models, one congregating *problematizing* and *ideating* into one factor (Model 1.3) and the other combining *problematizing* and *testing* (Model 1.4). As a result, both competing models showed poor fit. The fit statistics for Model 1.3 were: CFI = 0.912, TLI = 0.897, RMSEA (90% CI) = 0.072 (0.065, 0.079), SRMR = 0.047, and those for Model 1.4 were: CFI = 0.915, TLI = 0.901, RMSEA (90% CI) = 0.071 (0.063, 0.078), SRMR = 0.052. Again, discriminant validity was corroborated between the two pairs of factors under inspection.

Finally, we inspected the range of confidence interval of the correlation between the two pairs of factors in question. Discriminant validity would be established if the coverage did not cover the value of one (Anderson & Gerbing, 1988). The lower and upper limits of the confidence interval were calculated by adding to or subtracting from the correlation estimate two standard errors. For *problematizing* and *ideating*, the correlation estimate was 0.75, and the standard error was 0.05, the confidence interval thus ranging from 0.65 to 0.85. For *problematizing* and *testing*, the correlation estimate was 0.68, and the standard error was 0.04, the confidence interval thus ranging from 0.60 to 0.76. That both confidence intervals excluded the value of one suggested the two pairs of factors should not

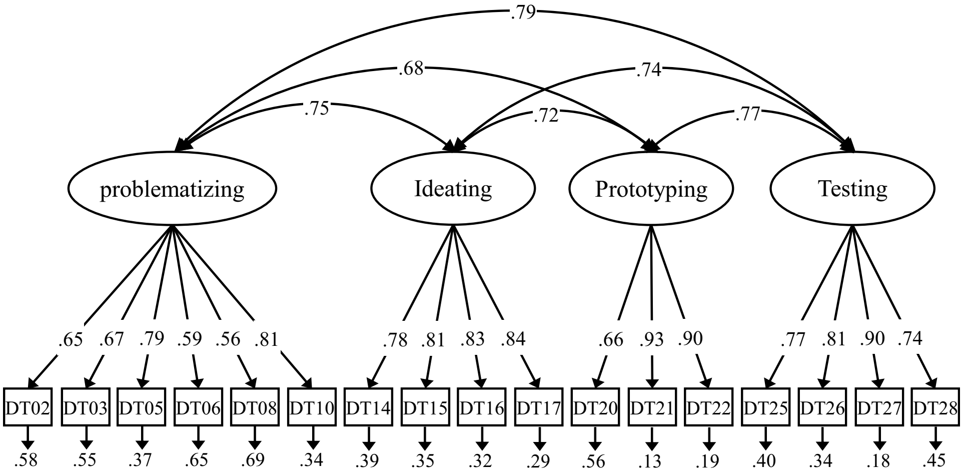

Fig. 1. Correlated four-factor model with standardized estimates ($n_2 = 509$).
All factor loadings are significant at $p < .001$.

Table 4
Statistics for reliability, convergent and discriminant validity ($n_2 = 509$).

						Squared correlation (r^2)			
Factor	α	ω	H	CR	AVE	Problematizing	Ideating	Prototyping	Testing
Problematizing	.88	.82	.86	.84	.47	–	.57	.46	.62
Ideating	.83	.88	.89	.88	.71		–	.52	.55
Prototyping	.81	.86	.94	.88	.71			–	.60
Testing	.89	.88	.90	.88	.66				–

Note. α = Cronbach's alpha; ω = omega; H = Coefficient H; CR = composite reliability; AVE = average variance explained.

belong to the same factor and further confirmed our assumption of discriminant validity.

MI of the four-factorial structure was assessed with the whole sample across subject (Mathematics and English) and gender. Table 5 shows the model-data fit results of the MI assessment. The analyses were done by placing a series of model constraints: configure invariance (Model 2.1 for subject groups and Model 3.1 for gender groups), weak invariance (Model 2.2 for subject groups and Model 3.2 for gender groups), and strong invariance (Model 2.3 for subject groups and Model 3.3 for gender groups) (Widaman & Reise, 1997). The model fits of all tested models were all at the acceptable level. From the simple model to the complex model, the changes in CFI values were smaller than the cut-off value of 0.01 (Cheung & Rensvold, 2002) and the changes in RMSEA values were smaller than 0.015 (Chen, 2007), suggesting strong measurement invariance of the SDTT across subject and gender groups.

4.4. Predictive validity

The resultant 17-item SDTT and the whole sample data were utilized in this step. First, as shown in Table 6, the bivariate correlation results indicated that all design thinking dimensions were positively correlated with instruction quality. The correlations ranged from 0.63 (with problematizing) to 0.63 (with testing), all with $p < .01$.

Second, a first-round CFA with the 5-item IQS produced a poor fit: CFI = 0.919, TLI = 0.838, RMSEA (90% CI) = 0.098 (0.076, 0.123), SRMR = 0.046. The modification indices suggested the associations between Item 1 and Item 2. After inspecting these items,

Table 5
Model fit indices ($N = 1018$).

Model	S-B χ^2	*df*	SRMR	RMSEA (90% CI)	Δ RMSEA	TLI	CFI	ΔCFI
Model 2.1	612.964*	226	.039	.058 (0.053, 0.064)	–	.932	.944	–
Model 2.2	634.373*	239	.054	.057 (0.052, 0.062)	.001	.934	.942	.002
Model 2.3	668.674*	252	.058	.057 (0.052, 0.062)	0	.934	.939	.003
Model 3.1	422.199*	226	.045	.064	–	.927	.939	–
Model 3.2	434.041*	239	.066	.062	.002	.931	.940	.001
Model 3.3	460.990*	252	.072	.062	0	.930	.935	.005

* $p < .001$; S-B χ^2 = Satorra–Bentler chi-square; df = degrees of freedom; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; TLI = Tucker–Lewis index; CFI = comparative fit index.

we found that both Item 1 and Item 2 were closely related to the classroom setting. Therefore, we modified the model by allowing the covariances between these items and this final model produced a good fit: CFI = 0.999, TLI = 0.997, RMSEA (90% CI) = 0.013 (0.000, 0.051), and SRMR = 0.012. Results of CFA with the 17-item SDTT indicated an acceptable fit: CFI = 0.942, TLI = 0.930, RMSEA (90% CI) = 0.058 (0.053, 0.063), and SRMR = 0.037.

Third, the MIMIC model included four design thinking factors as the exogenous variables and the instruction quality factor as the endogenous variable, with age, gender, subject, school level, education background, and year of teaching experience as covariates to control for possible confounding effects. The model showed an acceptable fit: CFI = 0.927, TLI = 0.911, RMSEA (90% CI) = 0.049 (0.046, 0.052), and SRMR = 0.039.

Fig. 2 displays the main paths with standardized estimates of the MIMIC model, and more specific information on path coefficients is shown in Supplementary Table 1. As shown, instruction quality was significantly predicted by *problematizing* ($\beta = 0.38$, $p < .001$) and *testing* ($\beta = 0.34$, $p < .001$). The effect of *prototyping* on instruction quality was positive but non-significant ($\beta = 0.22$, $p = .055$). The effect of ideating on instruction quality was null. Moreover, all four design thinking factors were positively related to each other, with *r*s ranging from 0.73 (between *problematizing* and *prototyping*) to 0.89 (between *ideating* and *prototyping*).

5. Discussion

This study aimed to develop and validate the SDTT to measure teachers' design thinking. The original pool of candidate items contained 29 items falling into five dimensions corresponding to the five-mode Stanford model: *empathizing, defining, ideating, prototyping*, and *testing*. Through cross-validating the scale with 1018 in-service teachers from 127 schools, the validated SDTT retained 17 items which fell into four subscales: *problematizing* (combining *emphasizing* and *defining*), *ideating, prototyping*, and *testing*. The scale showed good measurement validity, reliability, convergent validity, discriminant validity, and consequential validity. MIMIC results demonstrated that SDTT displayed good validity in predicting instruction quality mainly through *problematizing* and *testing* factors.

Among the four SDTT factors validated in our study, *ideating, prototyping*, and *testing* corresponded to the same labels in the Stanford model. The *problematizing* factor was the congregation of *empathizing* and *defining*. Our confirmation of the *ideating, prototyping*, and *testing* suggested that these factors also underlie design thinking in the teaching context, as indicated in previous literature (e.g., Atchia, 2021; Henriksen et al., 2018; Scheer et al., 2012). *Ideating* focuses on the creativity to generate teaching ideas through brainstorming, self-reflection, collaboration with colleagues, or vicarious experience of quality teaching. *Prototyping* emphasizes the action of teaching plan-making; during this course, teachers present the teaching process in a more visualized way, recognize the purposes of plan-making by grasping a deeper understanding of teaching methods, predicting potential problems in actual classroom teaching, and making revisions to address the problems. *Testing* is associated with the tryout of teaching plans in the classroom, during which teachers pay attention to evaluate the teaching plan based on self-reflection and students' responses to aid further improvement.

The *problematizing* factor demonstrated teachers' efforts in specifying the objectives or problems they must address. The combination of *empathizing* and *defining* has been suggested in the design thinking literature. According to Norman (2013), the Double-Diamond Design Process Model introduced by the British Design Council in 2005 divided design thinking into four processes: *discover, define, develop*, and *deliver*. In this model, *discover* and *define* are labeled as "finding the right problem" (p. 220). Brown (2008) posited design thinking embraced three spaces, i.e., *inspiration, ideation*, and *implementation*. Among them, the *inspiration* space consists of discovering peoples' needs and constructing the mental constraints that illustrate the starting point, criteria for evaluating progress, and objectives (Brown & Wyatt, 2010). In this sense, the inspiration space signifies a composite of *empathizing* and *defining*.

We contend that at least three reasons should have led to the aggregation of *empathizing* and *defining* in our study. First, to understand the problem, teachers could intentionally approach students and uptake students' perspectives through observation or communication. In this sense, empathizing with students and defining problems will likely be undertaken concurrently. The second reason should relate to the high uniformity of the K-12 education system, where teachers must strictly conform to the pedagogical standards and teaching objectives stipulated by the school, local, or even national curriculum. Take China as an example. The nine-year compulsory education curriculum standards specify not only the pedagogical principles and goals for the teachers but also detailed descriptions of expected learning outcomes and directions for evaluating these learning outcomes (e.g., Chinese Ministry of Education, 2022a, 2022b). Such top-down compulsory requirements on teachers and students should have led to the aggregation of *empathizing* and *defining*. Furthermore, there is a consensus that design thinkers are more open-minded and more tolerant of ambiguity (Blizzard et al., 2015; Brown, 2008; Owen, 2007). Most problems that teachers strive to solve are caused by various sources, such as teaching content, students' attainment levels, school atmosphere, communication with parents, and so forth (Norton & Hathaway, 2015). Teachers with design thinking embrace uncertainty and possibilities that arise in the *empathizing* and *defining* situations (Henriksen et al., 2018). This should have contributed to the highly intertwined performance of *empathizing* and *defining*.

Table 6
Correlations ($N = 1018$).

Variables	Problematizing	Ideating	Prototyping	Testing
Instruction quality	.63	.59	.57	.63
Problematizing	–	.69	.61	.67
Ideating	–	–	.73	.69
Prototyping	–	–	–	.70

Note. All correlations are significant at $p < .01$.

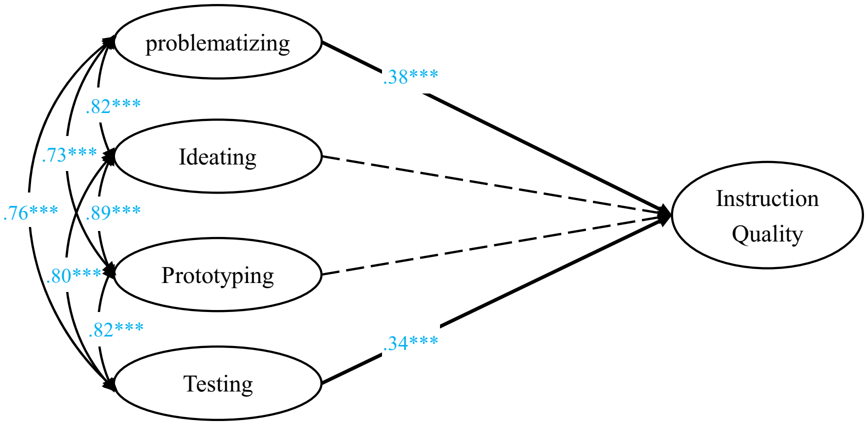

Fig. 2. MIMIC model results with standardized estimates (N = 1018).
*p < .05; **p < .01; ***p < .001; Dashed lines indicate non-significant paths.
For the sake of clarity, covariates are omitted. Standardized estimates for the covariates are displayed in Supplementary Table 1.

As for the predictive validity, results showed that out of the four design thinking factors, *problematizing* and *testing* directly predicted instruction quality positively, while the other two factors did not show direct and significant effects on instruction quality.

The positive effect of *problematizing* on instruction quality is in line with previous studies. Our results suggested that teachers of higher *problematizing* emphasize giving feedback activating students' higher-order thinking, and evaluating learning outcomes using different strategies. Note that an essential component of *problematizing is teacher empathy,* a positive teacher attribute that helps teachers to empower students to reduce learning obstacles and facilitate the learning process (Meyers et al., 2019). *A teacher with high empathy* privileges students' subjective feelings and is attentive to what students need in learning (Hen & Sharabi-Nov, 2014). An empirical study has documented that teacher empathy was positively associated with effective problem-solving skills and competence to handle students' problem behaviors (Wink et al., 2021). Further, *problematizing* reflects teachers' constant clarification of instructional difficulties and characteristics of quality teaching. Teachers constantly reviewing and articulating the essentials of teaching are more likely to present enhanced clarity of instruction.

Like *problematizing, testing* also showed a significant positive effect on instruction quality. This significant effect of *testing* on instruction quality can be attributed to three reasons. First, *testing* is directly related to teaching practice among the four factors. It manifests the direct implementation of teaching plans and involves actual instruction. Second, *testing* requires teachers to monitor the process of delivering the teaching plan and evaluate the efficiency and effectiveness of instruction. In this sense, *testing* involves metacognitive teaching, a key component of effective instruction extensively documented in relevant research (Jiang et al., 2016). Third, as instruction quality in the present study was reported by teachers, it can be interpreted as an indicator of self-belief about their instruction competence, namely, self-efficacy, self-concept, or self-confidence. For instance, Voogt et al. (2011) reviewed nine studies on teacher learning and concluded that implementing *testing* improved teachers' self-confidence in teaching. Teachers' engagement in testing the teaching plans could help them recognize how to improve and optimize their teaching process. Besides, positive responses from students are likely to elevate teachers' self-belief, and negative feedback from students can be helpful for teachers to undertake further refinement in their instruction.

The two factors *ideating* and *prototyping* failed to predict instruction quality significantly. However, this should not be interpreted as they did not influence instruction quality. Recall that in our earlier correlation analysis, all four factors were significantly related to instruction quality. The most important reason for the non-significant relations of *ideating* and *prototyping* to instruction quality should be that these factors indirectly exerted their influence through *problematizing* and *testing*. These inferences can be seen from the significant associations between the design thinking factors.

One possibility is that *testing* might mediate the relation between *prototyping* and instruction quality. This is because when teachers prototype, they usually generate and organize new ideas into a workable teaching plan, try to visualize it with flow charts, or illustrate the plan in concrete terms with examples. With more time and effort devoted to *prototyping*, teachers might become more inclined to test the teaching plan. In this way, *testing* is likely to pass over the effect of *prototyping* on instruction quality.

Besides, *ideating* might indirectly influence instruction quality through *prototyping* or *testing*. First, it is natural for teachers to follow the workflow where ideas are used to design a teaching plan, which is subsequently tested in the classroom. It is reasonable to surmise that the indirect effect of ideating on instruction quality be transferred through *prototyping* or *testing*. Second, design thinking has been characterized as a nonlinear process during which design thinkers do not strictly follow a specific sequence but constantly move back and forth or even skip over one or more steps (Leifer & Steinert, 2011). For instance, when a new teaching idea arises, teachers may directly test this idea in classroom instruction. In this way, *testing* may serve as the mediator to pass the effect of *ideating* on instruction quality.

6. Conclusion

The current study aimed to develop and validate the SDTT to measure design thinking for teachers. The factorial structure of the SDTT was cross-validated based on the five-mode Stanford Design Thinking Model, and the scale's reliability, convergent validity, and discriminant validity were established. Results demonstrated that the SDTT was a reliable scale consisting of four dimensions: *problematizing, ideating, prototyping*, and *testing*, with strong measurement invariance across subject (Mathematics and English) and gender groups. In terms of predictive validity, instruction quality was significantly predicted by *problematizing* and *testing* factors. Overall, we concluded that the SDTT confirmed the Stanford Design Thinking Model and that design thinking was a good predictor of teachers' instruction quality.

7. Limitations and implications

The current project had several limitations. First, the target population was mathematics and EFL teachers at all grades in the K-12 education system. However, due to resource constraints, we were only able to recruit mathematics and EFL teachers teaching at the primary school (K-6) and junior middle school (K-9) levels, leaving teachers at the senior middle school level (K-10 to K-12) unattended. Future studies are encouraged to include teachers from the entire span of the K-12 education system. Second, instruction quality was measured using a self-reported instrument, and this subjective measure of teaching efficiency is at risk of bias due to social desirability. Future studies may consider using external measurements such as student ratings, supervisor ratings of instruction quality, or students' English achievement scores. The last limitation to mention regards the cross-sectional nature of the current project. This static exploration leaves uncertainty regarding the temporal stability of the SDTT, but also the assumed indirect effects of *ideating* and *prototyping* on instruction quality. Future studies should consider a longitudinal design to explore the possible indirect effects of *prototyping* and *ideating* on instruction quality.

Nonetheless, our project can advance research and practice of design thinking for EFL and Mathematics teaching in at least two ways. The most salient contribution resides in that our study provided a reliable and valid instrument for measuring designing thinking for teaching in Mathematics and EFL and teaching in other relevant domains. The notion of design thinking and its importance to instruction quality has long been embraced in the relevant literature. However, the field still lacks a reliable and valid instrument to measure teachers' design thinking. Due to this absence, existing studies are usually conducted on small scales and with a small number of teachers. Therefore, the findings of these studies inevitably bear idiosyncrasies that are hard to generalize to a larger population. Our development of the SDTT is timely and should be able to address the vacancy of a reliable and valid measure for teachers' design thinking. This availability would facilitate design thinking researchers to obtain large-scale quantitative data more conveniently and more efficiently. Drawing on this measure, future researchers could conveniently examine the relation of design thinking to instructional quality, identify what factors can enhance design thinking, and explore the conditions in which design thinking functions most efficiently in enhancing instructional quality. All these studies could generate information that would help improve teacher training effectiveness. Pedagogically, the SDTT items can be regarded as concrete and operable strategies for teachers to apply in classroom teaching. For instance, teachers can use the scale items to examine their students' needs, identify the problems of instruction, develop creative teaching plans, deliver their instruction according to their plans, and iteratively evaluates the whole process of designing their instruction activities. Besides, the scale can also be used by teacher educators to diagnose the weaknesses and strengths of pre-or in-service teachers in teacher training programs.

CRediT authorship contribution statement

Yuyang Cai: Conceptualization, Methodology, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Yan Yang:** Data curation, Formal analysis, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors have no competing interests to declare that are relevant to the content of this article.

Data availability

The data that has been used is confidential.

Funding

This work was supported by The Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning [Code: TP2018068] given to the first author.

Acknowledgments

We thank Dr Bei Yuan and Zhongshan Teacher Development Center for their assistance with the data collection. Our thanks also go

to teachers for their time and participation in the study.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.tsc.2023.101255.

References

Aichholzer, J. (2014). Random intercept EFA of personality scales. *Journal of Research in Personality, 53*, 1–4. https://doi.org/10.1016/j.jrp.2014.07.001

Ainley, J., & Carstens, R. (2018). Teaching and learning international survey (TALIS) 2018 conceptual framework. *OECD Education Working Papers, No. 187*. OECD Publishing. https://doi.org/10.1787/799337c2-en

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103*(3), 411. https://doi.org/10.1037/0033-2909.103.3.411

Atchia, S. M. (2021). Integration of 'design thinking' in a reflection model to enhance the teaching of biology. *Journal of Biological Education*, 1–15. https://doi.org/10.1080/00219266.2021.1909642

Bento, C., Pereira, A. T., Azevedo, J., Saraiva, J., Flett, G. L., Hewitt, P. L., et al. (2019). Development and validation of a short form of the child–adolescent perfectionism Scale. *Journal of Psychoeducational Assessment, 38*(1), 26–36. https://doi.org/10.1177/0734282919879834

Bernardo, A. B. I., Cai, Y., & King, R. B. (2021). Society-level social axiom moderates the association between growth mindset and achievement across cultures. *British Journal of Educational Psychology, 91*(4), e12411. https://doi.org/10.1111/bjep.12411

Blizzard, J., Klotz, L., Potvin, G., Hazari, Z., Cribbs, J., & Godwin, A. (2015). Using survey questions to identify and learn more about those who exhibit design thinking traits. *Design Studies, 38*, 92–110. https://doi.org/10.1016/j.destud.2015.02.002

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). Guilford Press.

Brown, T. (2008). Design thinking. *Harvard Business Review, 86*(6), 84. Retrieved October 9, 2021, from https://hbr.org/2008/06/design-thinking.

Brown, T., & Wyatt, J. (2010). Design thinking for social innovation. *Stanford Social Innovation Review, winter*, 30–35. https://doi.org/10.1596/1020-797X_12_1_29

Buchanan, R. (1992). Wicked problems in design thinking. *Design issues, 8*(2), 5–21. https://doi.org/10.2307/1511637

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Chinese Ministry of Education. (2022a). *English curriculum standards for compulsory education*. Beijing Normal University Publishing Group.

Chinese Ministry of Education. (2022b). *Mathematics curriculum standards for compulsory education*. Beijing Normal University Publishing Group.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98–104. https://doi.org/10.1037/0021-9010.78.1.98

Crites, K., & Rye, E. (2020). Innovating language curriculum design through design thinking: A case study of a blended learning course at a Colombian university. *System, 94*. https://doi.org/10.1016/j.system.2020.102334

Cross, N. (2001). Designerly ways of knowing: Design discipline versus design science. *Design Issues, 17*(3), 49–55. Retrieved March 2, 2021, from http://www.jstor.org/stable/1511801.

Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measures: consistency, stability, and convergent validity. *Psychological Science, 12*(2), 163–170. https://doi.org/10.1111/1467-9280.00328

National People's Congress. (2006). Compulsory education law of the people's republic of China. Retrieved April 16, 2022, from http://en.moe.gov.cn/documents/laws_policies/201506/t20150626_191391.html.

d.school. (2018). Design thinking bootleg. Retrieved November 16, 2020, from https://dschool.stanford.edu/resources/design-thinking-bootleg.

Dorst, K. (2011). The core of 'design thinking' and its application. *Design Studies, 32*(6), 521–532. https://doi.org/10.1016/j.destud.2011.07.006

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*(3), 399–412. https://doi.org/10.1111/bjop.12046

Ellis, R. (2017). Position paper: Moving task-based language teaching forward. *Language Teaching, 50*(4), 507–526. https://doi.org/10.1017/S0261444817000179

English, A. R. (2016). John dewey and the role of the teacher in a globalized world: Imagination, empathy, and 'third voice. *Educational Philosophy and Theory, 48*(10), 1046–1064. https://doi.org/10.1080/00131857.2016.1202806

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*(1), 39–50. https://doi.org/10.1177/002224378101800104

Goi, H. C., & Tan, W. L. (2021). Design thinking as a means of citizen science for social innovation. *Front Sociology, 6*, Article 629808. https://doi.org/10.3389/fsoc.2021.629808

Goldman, S., Kabayadondo, Z., Goldman, S., & Kabayadondo, Z. (2017). Taking design thinking to school: How the technology of design can transform teachers, learners, and classrooms. *Taking design thinking to school: How the technology of design can transform teachers, learners, and classrooms* (pp. 3–19). Routledge.

Grisay, A. (2003). Translation procedures in OECD/PISA 2000 international assessment. *Language Testing, 20*(2), 225–240. https://doi.org/10.1191/0265532203lt254oa

Grisay, A. (2006). Translation and cultural appropriateness of the test and survey material. *PISA 2003 Technical Report*. https://doi.org/10.1787/9789264010543-6-en

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). Prentice Hall.

Hayton, J. C., Allen, D. G., & Scarpello, V. (2016). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods, 7*(2), 191–205. https://doi.org/10.1177/1094428104263675

Hen, M., & Sharabi-Nov, A. (2014). Teaching the teachers: Emotional intelligence training for teachers. *Teaching Education, 25*(4), 375–390. https://doi.org/10.1080/10476210.2014.908838

Henriksen, D., Gretter, S., & Richardson, C. (2018). Design thinking and the practicing teacher: Addressing problems of practice in teacher education. *Teaching Education, 31*(2), 209–229. https://doi.org/10.1080/10476210.2018.1531841

Hmelo-Silver, C. E. (2004). Problem-based learning: What and how do students learn? *Educational Psychology Review, 16*(3), 235–266. https://doi.org/10.1023/B:EDPR.0000034022.16470.f3

Jaber, L. Z., Southerland, S., & Dake, F. (2018). Cultivating epistemic empathy in preservice teacher education. *Teaching and Teacher Education, 72*, 13–23. https://doi.org/10.1016/j.tate.2018.02.009

Jiang, Y., Ma, L., & Gao, L. (2016). Assessing teachers' metacognition in teaching: The teacher metacognition inventory. *Teaching and Teacher Education, 59*, 403–413. https://doi.org/10.1016/j.tate.2016.07.014

Kenny D.A. (2016). Multiple factor models: Confirmatory factor analysis. Retrieved June 16, 2022, from http://davidakenny.net/cm/mfactor.htm#DV.

Khodabakhshzadeh, H., Hosseinnia, M., Moghadam, H. A., & Ahmadi, F. (2018). EFL teachers' creativity and their teaching's effectiveness: A structural equation modelling approach. *International Journal of Instruction, 11*(1), 227–238. https://doi.org/10.12973/iji.2018.11116a

Kim, Y. E., Brady, A. C., & Wolters, C. A. (2018). Development and validation of the brief regulation of motivation scale. *Learning and Individual Differences, 67*, 259–265. https://doi.org/10.1016/j.lindif.2017.12.010

Kline, R. B. (2015). *Principles and practice of structural equation modeling.* Guilford publications.

Koh, J. H. L., Chai, C. S., Wong, B., & Hong, H. Y. (2015). *Design thinking for education: Conceptions and applications in teaching and learning.* Singapore: Springer. https://doi.org/10.1007/978-981-287-444-3

Krosnick, J. A., Holbrook, A. L., Berent, M. K., Carson, R. T., Michael Hanemann, W., Kopp, R. J., et al. (2002). The impact of "No opinion" response options on data quality: Non-attitude reduction or an invitation to satisfice? *Public Opinion Quarterly, 66*(3), 371–403. https://doi.org/10.1086/341394

Leifer, L. J., & Steinert, M. (2011). Dancing with ambiguity: Causality behavior, design thinking, and triple-loop-learning. *Information Knowledge Systems Management, 10*(1–4), 151–173. https://doi.org/10.3233/IKS-2012-0191

Liedtka, J. (2015). Perspective: Linking design thinking with innovation outcomes through cognitive bias reduction. *Journal of Product Innovation Management, 32*(6), 925–938. https://doi.org/10.1111/jpim.12163

Lin, L., Shadiev, R., Hwang, W. Y., & Shen, S. (2020). From knowledge and skills to digital works: An application of design thinking in the information technology course. *Thinking Skills and Creativity, 36*. https://doi.org/10.1016/j.tsc.2020.100646

Long, M. (2014). *Second language acquisition and task-based language teaching.* Wiley-Blackwell.

Madson, M. J. (2021). Making sense of design thinking: A primer for medical teachers. *Medical Teacher, 43*(10), 1115–1121. https://doi.org/10.1080/0142159X.2021.1874327

Matsui, H. (2020). Design thinking for transforming a foreign language curriculum: From traditional curriculum to personalized flipped curriculum. In *Proceedings of the EdMedia + Innovate Learning.* Retrieved May 28, 2022, from https://www.learntechlib.org/p/217343.

McLaughlin, J. E., Wolcott, M. D., Hubbard, D., Umstead, K., & Rider, T. R. (2019). A qualitative review of the design thinking framework in health professions education. *BMC Medical Education, 19*(1), 98. https://doi.org/10.1186/s12909-019-1528-8

Meyers, S., Rowell, K., Wells, M., & Smith, B. C. (2019). Teacher empathy: A model of empathy for teaching for student success. *College Teaching, 67*(3), 160–168. https://doi.org/10.1080/87567555.2019.1579699

Motallebzadeh, K., Ahmadi, F., & Hosseinnia, M. (2018). The relationship between EFL teachers' reflective practices and their teaching effectiveness: A structural equation modeling approach. *Cogent Psychology, 5*(1), Article 1424682. https://doi.org/10.1080/23311908.2018.1424682

Mueller, R. O., Hancock, G. R., Hancock, G. R., Stapleton, L. M., & Mueller, R. O. (2019). Structural equation modeling. *The reviewer's guide to quantitative methods in the social sciences* (pp. 445–456). Routledge, 2nd ed.

Muthén, L. K., & Muthén, B. Q. (2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.

Muthén, L. K., & Muthén, B. Q. (2020). *Mplus 8.5 [computer software].* Muthén & Muthén.

Nation, I. P., & Macalister, J. (2019). *Language curriculum design.* Routledge.

Norman, D. (2013). *The design of everyday things: Revised and expanded edition.* Basic books.

Norton, P., & Hathaway, D. (2015). In search of a teacher education curriculum: Appropriating a design lens to solve problems of practice. *Educational Technology*, 3–14. Retrieved December 2, 2020, from https://www.jstor.org/stable/44430419.

OECD. (2018). Teaching and learning international survey (TALIS): Teacher questionnaire. Retrieved September 25, 2021, from http://www.oecd.org/education/school/TALIS-2018-MS-Teacher-Questionnaire-ENG.pdf.

OECD. (2019). *PISA 2018 assessment and analytical framework*. OECD publishing. https://doi.org/10.1037/0012-1649.28.5.759

Owen, C. (2007). Design thinking: Notes on its nature and use. *Design Research Quarterly, 2*(1), 16–27. Retrieved January 1, 2021, from https://www.id.iit.edu/wp-content/uploads/2015/03/Design-thinking-notes-on-its-nature-and-use-owen_desthink071.pdf.

Radford University. (2013). *Design thinking for educators-A radford university methodology workbook*. Retrieved May 16, 2021, from https://www.slideshare.net/ccvidadmin/design-thinking-for-ed-wbk1c-for-screen-26174639.

Razzouk, R., & Shute, V. (2012). What is design thinking and why is it important? *Review of Educational Research, 82*(3), 330–348. https://doi.org/10.3102/0034654312457429

Robertson, A. D., Atkins, L. J., Levin, D. M., Richards, J., Robertson, A. D., Atkins, L. J., Levin, D. M., & Richards, J. (2015). What is responsive teaching?. *Responsive teaching in science and mathematics* (1st ed, pp. 1–35). Routledge.

Roth, K., Globocnik, D., Rau, C., & Neyer, A. K. (2020). Living up to the expectations: The effect of design thinking on project success. *Creativity and Innovation Management, 29*(4), 667–684. https://doi.org/10.1111/caim.12408

Sarooghi, H., Sunny, S., Hornsby, J., & Fernhaber, S. (2019). Design thinking and entrepreneurship education: Where are we, and what are the possibilities? *Journal of Small Business Management, 57*(S1), 78–93. https://doi.org/10.1111/jsbm.12541

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66*(4), 507–514. https://doi.org/10.1007/BF02296192

Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika, 75*(2), 243–248. https://doi.org/10.1007/s11336-009-9135-y

Scheer, A., Noweski, C., & Meinel, C. (2012). Transforming constructivist learning into action: Design thinking in education. *Design and Technology Education: An International Journal, 17*(3). https://ojs.lboro.ac.uk/DATE/article/download/1758/1648.

Tsai, M. J., & Wang, C. Y. (2020). Assessing young students' design thinking disposition and its relationship with computer programming self-efficacy. *Journal of Educational Computing Research, 59*(3), 410–428. https://doi.org/10.1177/0735633120967326

Tseng, J. J., Cheng, Y. S., & Yeh, H. N. (2019). How pre-service English teachers enact TPACK in the context of web-conferencing teaching: A design thinking approach. *Computers and Education, 128*, 171–182. https://doi.org/10.1016/j.compedu.2018.09.022

Van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European journal of developmental psychology, 9*(4), 486–492. https://doi.org/10.1080/17405629.2012.686740

Voogt, J., Westbroek, H., Handelzalts, A., Walraven, A., McKenney, S., Pieters, J., et al. (2011). Teacher learning in collaborative curriculum design. *Teaching and Teacher Education, 27*(8), 1235–1244. https://doi.org/10.1016/j.tate.2011.07.003

Walker, A., Lee, M., & Bryant, D. A. (2015). Development and validation of the international baccalaureate learner profile questionnaire (IBLPQ). *Educational Psychology, 36*(10), 1845–1867. https://doi.org/10.1080/01443410.2015.1045837

Widaman, K. F., Reise, S. P., Bryant, K. J., Windle, M., & West, S. G. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). American Psychological Association.

Wink, M. N., LaRusso, M. D., & Smith, R. L. (2021). Teacher empathy and students with problem behaviors: Examining teachers' perceptions, responses, relationships, and burnout. *Psychology in the Schools, 58*, 1575–1596. https://doi.org/10.1002/pits.22516

Wu, B., Hu, Y., & Wang, M. (2019). Scaffolding design thinking in online STEM preservice teacher training. *British Journal of Educational Technology, 50*(5), 2271–2287. https://doi.org/10.1111/bjet.12873

Yoon, M., & Lai, M. H. C. (2018). Testing factorial invariance with unbalanced samples. *Structural Equation Modeling: A Multidisciplinary Journal, 25*(2), 201–213. https://doi.org/10.1080/10705511.2017.1387859

Zeichner, K. M., & Liston, D. P. (2013). *Reflective teaching: An introduction* (2nd ed.). Routledge. https://doi.org/10.4324/9780203771136

IDEO. (2016). Design thinking for educators toolkit (2nd ed.) Retrieved November 18, 2020, from https://www.ideo.com/post/design-thinking-for-educators.