CROSS-CULTURAL DIFFERENCES IN DIVERGENT THINKING AND EVALUATIVE SKILL

by

YAWEI GUO

(Under the Direction of Tarek Grantham and Bonnie Cramond)

ABSTRACT

As an important component of creativity, evaluative skill refers to the ability to accurately judge ideas on creativity or originality. Although the importance of the evaluative component is recognized in most creativity models, studies on this construct are lacking compared to those on divergent thinking (DT). Creativity research in the cross-cultural context is an emerging field. In light of the differences in the creativity conception that compared to Westerners (Americans and Europeans), Easterners (Asians) put more emphasis on following traditions and norms, this evaluative component should be important in Asians' creativity. Therefore, the present study investigated the differences in evaluative skill as well as DT between American (n = 341) and Chinese (n= 345) colleges students with four types of DT tests and corresponding evaluation tasks (Line Meanings, Uses, Instances, and Consequences) via latent mean comparisons, which were conducted after the measurement invariance (MI) of the measures was established. Results supported the multidimensionality of both constructs of DT and evaluative skill based on the different tasks used. Multi-group confirmatory factor analyses supported configural, weak and strong MI for the revised multidimensional model for both DT and evaluative skill. In addition, latent mean comparisons demonstrated higher performance of

American individuals on DT fluency (on Line Meanings, Uses and Consequences) and DT originality (on Line Meanings and Consequences) compared to their Chinese counterparts. It also showed higher performance of American respondents on evaluative skill based on the Uses evaluation task compared to Chinese peers, whereas no difference emerged on that based on the Line Meanings evaluation task. Last, in terms of the cultural orientation used to interpret the differences, structural equation modelling results showed that the dimension of vertical collectivism consistently negatively predicted both DT and evaluative skill across different tasks. Overall, the findings suggested that there are cross-cultural differences in people's evaluative skill, and this pattern is different from that in divergent thinking. This study also demonstrates the importance of testing MI of the measures before comparing creativity differences cross-culturally.

INDEX WORDS: Creativity, Divergent Thinking, Evaluative Skill, Cross-Cultural, Cultural Orientation, Individualism, Collectivism, Measurement Invariance

CROSS-CULTURAL DIFFERENCES IN DIVERGENT THINKING AND EVALUATIVE SKILL

by

YAWEI GUO

B. A., Nanchang University, China, 2011

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2021

CROSS-CULTURAL DIFFERENCES IN DIVERGENT THINKING AND

EVALUATIVE SKILL

by

YAWEI GUO

Major Professor:	Tarek C. Grantham
Co-major Professor:	Bonnie Cramond
Committee:	Amanda Ferster
	Daehyun Kim
	Mark A. Runco

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
December 2021

ACKNOWLEDGEMENTS

I would like to acknowledge my major professors Dr. Tarek Grantham and Dr. Bonnie Cramond for their persistent support and helpful insights during my whole PhD study as well as this dissertation process. For Dr. Grantham, I acknowledge his guidance in the whole dissertation process, and providing various opportunities to enhance teaching skills. For Dr. Cramond, I appreciate her welcoming me to The University of Georgia and guidance in doing research. Besides, I would like to express my appreciation to my committee members: Dr. Amanda Ferster, Dr. Daehyun Kim, and Dr. Mark Runco. Specifically, Dr. Ferster has provided invaluable insights and resources on statistical analysis strategies. Though I felt challenged by the advanced methods initially, I benefited a lot from this learning process. For Dr. Kim, I appreciate her expertise and insights on culture-relevant issues, and for Dr. Runco, I benefited from his guidance on the research topic and reliability issues.

In addition, my special thanks go to my colleagues Shengjie Lin, who provided insightful comments for refining this dissertation, and Zack Williams, who provided substantive suggestions on the data analysis, as well as the proofreading. Further, I appreciate the assistance and suggestions of my other colleagues and teachers in the processes of data collection and data analysis, including Lili Clark, Xizheng Xu, Chenyu Shangguan, Shaoying Gong, Yang Cao, Qichen Zhao, Jiajun Guo, Lesa Hoffman, and Laura Lu. The completion of the dissertation would not be possible without their support. Last, I would like to extend my appreciation to all who supported me and participated in my dissertation study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

This chapter begins by introducing the main concepts and stating the problems to be researched. Then it introduces the goals of the study, followed by the research questions, and the overview of the study.

History is replete with cases in which creative ideas or work were initially rejected, which brought great losses to individuals or even the society (Licuanan et al., 2007; Mueller et al., 2018). For example, J. K. Rowling's best seller *Harry Potter* was initially rejected by many publishers. Kodak failed to identify the creativity of digital photography, and thus, nearly put itself into bankruptcy (Lucas & Gog, 2009). In addition, the flying machine invented by the Wright Brothers was ignored by the US military at first sight (Licuanan et al., 2007).

The above-mentioned examples point to the problem in terms of creative idea evaluation. Evaluative skill (Runco & Smith, 1992), or the ability to accurately identify creative ideas, is recognized in most creativity theories (e.g., Campbell, 1960; Cropley, 2006; Reiter-Palmon et al., 2012; Simonton, 1988; Wallas, 1926), and is critical to creative achievement and creative idea implementation (Kasof, 1995; Runco & Smith, 1992). However, creativity research thus far has emphasized idea generation or divergent thinking (DT), and empirical research on the evaluative component is relatively lacking (Kozbelt, 2007; Silvia, 2008). According to Cropley (2006), poor evaluative skill can result in at least two negative outcomes. First, if people rate a novel idea as common or unoriginal, they may lose the opportunity, and thus, inhibit creativity. Second, if individuals judge a commonplace idea as creative this may bring the risk of seriously negative changes (e.g., a misallocation of resources to that idea). As mentioned above, there have been many cases in history in which creative ideas or products were rejected; therefore, advanced evaluative skill could benefit the society.

Creativity research in the cross-cultural context is an emerging field (Karwowski et al., 2020; Niu & Kaufman, 2013); however, similarly to creativity research more broadly, most studies in this area have focused on DT, and research on evaluative thinking is rare (Ivancovsky et al., 2019). When considering cross-cultural differences in the creativity concept between Westerners (Americans and Europeans) and Easterner (Asians), the latter group tends to put more emphasis on following traditions and norms even for new ideas. Therefore, this evaluative component should be important in Asians' creativity (Niu & Sternberg, 2006). In terms of the relationship between DT and evaluative skill, Charles and Runco (2001) hypothesized that the fourth-grade slump in children's DT may be related to an increase in their evaluative abilities. In a similar vein, it is assumed that differences observed between East Asians and Western populations on DT tests (e.g., Deng et al., 2016; Zha et al., 2006) may be relevant to their differences in evaluative skill. Therefore, the main goal of the present dissertation study was to investigate cross-cultural differences in evaluative skill.

Notably, when administering measures to people from different cultures, it is important to ensure the measures assess the construct in the same way, that is, the construct has the same meaning interpreted by different groups. In other words, measurement invariance (MI) across the groups should be satisfied (Byrne & Watkins, 2003). Different cultures may have different influences on individuals' perceptions and manifestations of creativity (Niu & Sternberg, 2002), and researchers have previously expressed concerns about the construct validity of creativity tests (e.g., DT tests) administered in cross-cultural contexts (e.g., Karwowski, 2016; Runco, 2004). Despite the concerns, studies testing MI in the creativity field are generally lacking (Karwowski, 2016; Kuhn & Holling, 2009), particularly in cross-cultural contexts. Thus, the present study aimed to investigate cross-cultural differences in evaluative skill based on measures with established MI to reach more statistically robust conclusions. Because previous cross-cultural comparison studies on DT did not address the issue of MI (Guo et al., 2021) either, the second goal of this study was to further investigate cross-cultural differences in DT based on measures with assumed MI.

In addition, regarding the reasons why there are cross-cultural differences in creativity, it is common for researchers to adopt cultural orientation, that is, individualism vs. collectivism, as the main influential factor (Ng & Smith, 2004). However, most studies have not explicitly measured the concept but may just regard nationality as a proxy (Xie & Paik, 2019). Also, as there are other dimensions of cultural orientation other than individualism vs. collectivism, for example uncertainty avoidance (Hofstede, 1980a; Jung, 2002), it is necessary to systematically investigate how various dimensions of cultural orientation affect creativity outcomes using direct measures of the concepts. Thus, the third goal of the present study was to investigate the relationships between cultural orientation and creativity based on measures found to have satisfactory MI.

Taken together, the main goal of the dissertation study was to investigate cross-cultural differences in evaluative skill (applying measures with MI satisfied), specifically, between American and Chinese individuals. Secondary objectives included further investigating cross-cultural differences in DT applying measures with MI established. Last, for all cross-cultural differences in evaluative skill and/or DT that were found, the study went on to explore whether multiple facets of cultural orientation (i.e., individualism vs. collectivism, and uncertainty avoidance) can predict evaluative skill and/or DT.

Research Questions

Among the following seven questions, Questions 1–4 (latent structure and MI tests) were applied to the measures for all constructs of interest including divergent thinking, evaluative skill, individualism-collectivism, and uncertainty avoidance.

For American and Chinese groups,

RQ1: What is the latent structure of the measures?

RQ2: Do the measures have the same factors and relationships between indicators and factors across the two cultural groups?

RQ3: For the measures, are factors linked with corresponding indicators to the same extent (strength) across the two cultural groups?

RQ4: For the measures, do American respondents obtain the same observed score on an indicator as Chinese peers who have the same ability on a respective latent factor?

RQ5: Are there differences between the two cultural groups in divergent thinking at the latent level (i.e., as a latent trait)?

RQ6: Are there differences between the two cultural groups in evaluative skill at the latent level (i.e., as a latent trait)?

RQ7: Is cultural orientation (individualism vs. collectivism, uncertainty avoidance) associated with DT and evaluative skill?

Overview of the Study

The aim of the dissertation study was to investigate cross-cultural differences in DT and evaluative skill, and the relationships between individualism vs. collectivism, uncertainty avoidance, DT and evaluative skill. College students from America and China were administered four types of DT tests and corresponding evaluation tasks (asking participants to rate ideas previously produced by others on DT tests): Line Meanings, Uses, Instances and Consequences. The cultural orientation measures (the individualism-collectivism test, the uncertainty avoidance test) and demographic measures were also administered. Multi-group confirmatory factor analyses (CFA) were conducted to investigate the latent factor structure and measurement invariance of the measures for each of the constructs of interest, based on which latent mean comparisons were conducted to examine the cross-cultural differences in DT as well as evaluative skill. In addition, the relationships between cultural orientation (individualism-collectivism, uncertainty avoidance), DT and evaluative skill were systematically investigated by general SEM models.

CHAPTER 2

LITERATURE REVIEW

This chapter first briefly reviews the theories on different components of creativity, including divergent thinking (DT) and evaluative skill. Next, it describes how the constructs of DT and evaluative skill are measured respectively. Then, it reviews the studies on the relationship between DT and evaluative skill, followed by a detailed literature review on culture and creativity. Last, the concept of measurement invariance (MI) is introduced and relevant MI studies on creativity are reviewed.

Different Components of Creativity

Creativity as a concept itself can be complex, but it is generally agreed that it represents the production of both novel and appropriate products or ideas (Plucker et al., 2004; Runco & Jaeger, 2012). Throughout the history of creativity research, many theories have emphasized the important role of both divergent thinking and evaluative skill in creativity (Campbell, 1960; Cropley, 2006; Reiter-Palmon et al., 2012; Runco & Chand, 1995; Wallas, 1926).

Divergent thinking (DT) is defined as “cognition that leads in various directions” (Runco, 1999, p. 577). DT deserves its place as one of the critical creative processes, because DT tests can provide people with unique information that is not offered by GPA or IQ tests (Acar & Runco, 2019). Specifically, DT tests can effectively predict creative performance in real life situations. For example, Plucker (1999, p. 103) reported that “just under half of the variance in adult creative achievement is explained by DT test scores, with the contribution of DT being more than three times that of intelligence quotients.”

Evaluative skill is defined as the ability to accurately assess each presented idea on important dimensions regarding creativity (e.g., originality, appropriateness, or creativity in general) (Grohman et al., 2006; Runco & Smith, 1992). The evaluative component was discussed in slightly different names in different models. For example, Wallas (1926) proposed four stages for creativity, which include preparation, incubation, illumination, and verification. The evaluative component is indicated in the stage of verification. Runco and colleagues (Runco & Chand, 1995; Runco & Vega, 1990) suggested a componential model of creative processes, which includes problem finding, ideation and evaluation. Similarly, in other models the evaluative activity was described in terms such as adoption criteria (Rodgers & Adhikarya, 1979), selective retention (Campbell, 1960; Simonton, 1999), or discernment (Silvia, 2008).

Thus far, researchers in the field of creativity have made remarkable progress in helping people understand idea generation (or divergent thinking); however, less is known about the other important component—idea evaluation (Basadur, 1995; Runco & Chand, 1994; Silvia, 2008). This can be a problem since idea evaluation is recognized in most models of creativity (Mumford et al., 1991; Reiter-Palmon et al., 2012; Simonton, 1988; Wallas, 1926), and its importance has been acknowledged in the process of creativity (Lonergan et al., 2004; Mumford, 2001), especially in creative achievement, and creative idea implementation (Kasof, 1995; Runco & Smith, 1992).

There are two possible reasons for the relative lack of research in evaluative skill. First, one may understand evaluative thinking the same as traditional concepts of critical thinking or convergent thinking. However, this may be a misconception because there is psychometric evidence from previous studies demonstrating the discriminant validity of those evaluation measures (Runco, 1991; Runco & Dow, 2004; Runco & Smith, 1992). Specifically, those studies showed nonsignificant correlations between the scores on convergent thinking or critical thinking and evaluative accuracy. Second, the reason may be relevant to the measures of evaluative skill. To measure evaluative performance, there should be a list of ideas produced first (Runco, 2020), which may complicate the design of the study. In addition, an evaluative accuracy index needs to be obtained from the raw scores of participants, and researchers had various views on the method of measuring evaluative accuracy (Benedek et al., 2016; Silvia, 2008). The measurement issues on evaluative skill will be discussed in detail in later sections.

Measurement of Divergent Thinking (DT)

Typically, DT tests require respondents to think of various ideas in response to an open-ended question or task (Acar et al., 2020). Popularly used DT tests include the Torrance Tests of Creative Thinking (TTCT; Torrance, 1966, 1974), the Guilford Tests (Christensen et al., 1953; Guilford, 1967, 1979), and the Wallach-Kogan Creativity Tests (WKCT; Wallach & Kogan, 1965). The TTCT figural tests are various picture completion tasks, and the TTCT verbal tests include Ask and Guess, Guess Causes, Guess Consequences, Product Improvement, Uses and Just Suppose; the Guilford tests include Consequences, Plot Titles and Uses; and WKCT cover a series of tasks, such as Uses, Instances, Similarities, Line Meanings and Pattern Meanings. Other types of DT tests have additionally been developed, such as problem solving and problem finding tasks (Okuda et al., 1991; Runco & Chand, 1994).

Different DT tests may vary in terms of their psychometric properties and are not equivalent (Acar et al., 2020; Runco et al., 2016). For example, studies (Okuda et al., 1991; Runco et al., 2016) showed that individuals obtained higher originality scores on some tasks (e.g., Titles) compared to others (e.g., real-world problem tests). It is often suggested that various types of DT tasks be used in research to reach conclusions that are more generalizable (Long &

Plucker, 2015; Runco et al., 2016). In addition to the type of tasks, individuals' DT outcomes may vary depending on the DT output. Specifically, to score DT tests, different DT outputs or indices can be obtained, including fluency, flexibility, originality, elaboration, etc. (Long & Plucker, 2015). Fluency refers to the total number of relevant ideas produced by participants on a certain DT task, which is the gatekeeper score for other indices; originality represents the number of original (or unique) ideas produced; flexibility refers to the number of categories of ideas; and elaboration represents the number of details added to ideas beyond conveying the basic idea. Moreover, the testing environment can have an impact on DT scores. For example, approaching a DT task with less strict time conditions in a game-like environment may promote DT performance (Paek et al., 2021; Said-Metwaly, et al., 2020a; Wallach & Kogan, 1965). In addition, the testing instructions are also relevant to DT outcomes. A recent meta-analysis (Acar et al., 2020) suggested that compared with the standard instructions, explicit instructions (i.e., asking participants to be creative) combined with the standard instructions (i.e., asking people to produce as many ideas as possible) promoted DT performance.

Measurement of Evaluative Skill

Typically, the measurement of evaluative skill consists of two stages: In Stage One, ideas (usually from DT tests) are produced, and they are scored in terms of originality or creativity; these scores can then serve as the standard (or criterion, right answer) afterwards; In Stage Two, respondents are required to rate the ideas presented in terms of originality or creativity (Grohman et al., 2006; Runco, 2020; Runco & Smith, 1992; Runco & Vega, 1990), and these ratings are compared to the criterion produced in Stage One. To operationalize evaluative skill, researchers typically use one of the five methods to obtain an evaluative accuracy index: Top choice, hit rates, difference from criterion, correlation with criterion, and integrating originality and appropriateness. (a) Top choice refers to that subjects choose the idea(s) they think are most creative, or their most favorite choice(s), and the ideas selected are compared to the criterion (Silvia, 2008). (b) In the hit rates method, subjects rate ideas on a Likert scale, and then their answers are compared to the predetermined criterion in terms of the number of correctly identified creative or normal ideas (Runco & Smith, 1992; Runco & Dow, 2004). (c) Difference from criterion refers to that participants' ratings (on a Likert scale) on creativity are subtracted from the standard, and the mean absolute difference obtained for each participant can be an indicator of the discrepancy from the right answer (Grohman et al., 2006). (d) In the correlation with criterion method, subjects' creativity ratings (on a Likert scale) are correlated with the actual number of people who proposed the idea, and the correlation coefficient score would be obtained for each person on each subtest (Charles & Runco, 2001; Runco, 1991; Runco & Vega, 1990). (e) In the method of integrating originality and appropriateness, the dimension of appropriateness is considered. Specifically, subjects assign ideas to three categories: creative ideas (high on both originality and appropriateness), common ideas (low on originality and high on appropriateness), or inappropriate ones (high on originality but low on appropriateness) (Benedek et al., 2016). It was found that people rarely rated ideas as low on both dimensions, so the researchers decided to keep those three categories. Regarding the strengths and weaknesses of each method (see details in Benedek eta al., 2016; Silvia, 2008), it is suggested that the first two methods (top choice, hit rates) may be easier to implement, but the first one is a relative judgement and the second one can be biased by participants' tendency to rate all ideas as creative or uncreative. The fifth method (integrating originality and appropriateness) has the strength of considering appropriateness, but subjects cannot rate the whole pool of initially generated ideas.

There are some factors that can affect the measure outcomes. (a) The different methods to obtain the evaluative accuracy index may have an impact on the outcomes as described above (Benedek et al., 2016; Silvia, 2008). (b) The difference in the type of tasks for idea evaluation (i.e., ideas rated may come from different types of DT tests) may also exert an influence (Runco & Dow, 2004). (c) Intrapersonal vs. interpersonal evaluation. Participants may have better evaluative performance in rating one's own ideas than others' since they understand their own ideational processes better (Runco & Smith, 1992). (d) Difference in evaluation instructions. For example, popularity instructions (i.e., asking participants to evaluate how many people among 10 can generate the idea) may increase evaluative performance compared to creativity instructions (i.e., asking them to rate the creativity of an idea on a Likert scale) because the former is easier for people to operationalize (Runco, 1991).

Divergent Thinking and Evaluative Skill

Using the various methods to measure evaluative skill mentioned above, a sizable body of research has examined the relationship between DT and evaluative skill. However, findings from this line of research have been largely inconsistent. On one hand, some supported a positive relationship. For example, Benedek et al. (2016) reported that evaluative skill was positively associated with DT as well as creative achievement in college students. Similarly, Runco (1991) found a positive link between DT and evaluative skill by investigating children. Runco and Vega (1990) also documented a positive relationship between the two by asking teachers and parents to assess ideas produced by children.

On the other hand, others indicated a negative or null relationship. For example, Grohman et al. (2006) found a negative link between DT originality and evaluative accuracy in high schoolers. Similarly, Guo et al. (2019) reported that participants with higher DT gave lower ratings to creative ideas compared to those with lower DT. In addition, Runco and Smith (1992) investigated university students and found that interpersonal evaluation was unrelated to DT. Charles and Runco (2001) found that there was null relation between DT fluency and evaluative accuracy in children.

In short, it is too early to be conclusive regarding the relationship between DT and evaluative accuracy, and more studies are needed to examine possible moderators of this relationship across studies. It should be noted that all findings of studies above were based on interpersonal evaluations, i.e., evaluating ideas produced by others. For intrapersonal evaluations (i.e., rating on one's own ideas), Grohman and colleagues (2006) showed that individuals with the highest overall DT scores were more accurate than those with lowest DT scores. Runco and Smith (1992) also reported a positive relation between DT and intrapersonal evaluative accuracy.

To interpret the inconsistencies of the findings, several different explanations have been proposed. A positive link between the two constructs has been theorized because both DT and evaluative skill are critical components of creativity (Cropley, 2006; Reiter-Palmon et al., 2012; Runco & Chand, 1995; Wallas, 1926). Specifically, Runco and Chand (1994) suggested that people with high DT skills excel in idea generation, and more familiar with the ideational process, thus they can more accurately evaluate ideas. Alternatively, other researchers have argued that these processes are independent of each other or even negatively correlated because DT belongs to the ideational process whereas evaluative skill is analytic (Grohman et al., 2006; Sternberg & Lubart, 1995). In addition, as mentioned above regarding the measurement of divergent thinking and evaluative skill, there are many factors that can affect the measure outcomes of both DT and evaluative skill, thus the inconsistencies of the findings may also be attributed to a combination of those factors, including the type of DT tasks, different DT indices, the method to obtain the evaluative accuracy index, the type of tasks for idea evaluation, the instructions used in the evaluation tasks, etc. Further, participant characteristics (e.g., age, gender) may also play a role in affecting the relationship between DT and evaluative skill.

Last, previous studies suggested that people from different cultures may demonstrate distinct creativity and evaluation patterns (Ivancovsky et al, 2019; Karwowski, 2016), and thus, culture can be another factor that leads to differences in divergent thinking, differences in evaluative skill, as well as differences in the relationship between divergent thinking and evaluative skill. Therefore, a detailed review on culture and creativity is provided in the following section.

Culture and Creativity

Creativity: East and West

Creativity as a concept can be complex and researchers have proposed different models. One of the most popular was the 4P model suggested by Rhodes (1961). Within this model, creativity can be understood from four aspects: person, process, product and press (environment). Individuals' creativity can be affected by the surrounding environment, and culture can be regarded as one of the environmental factors. Different cultures may have different values and norms, which may differentially influence various aspects of creativity (Kim, 2016; Niu & Kaufman, 2013; Niu & Sternberg, 2002). Most cross-cultural studies of creativity have been conducted between respondents from individualistic cultures such as the US and those from collectivistic cultures such as China and Japan (Niu & Kaufman, 2013; Puente-Diaz et al., 2016).

East Asians include Confucianism-influenced populations, for example, those from countries or regions of Mainland China, Hong Kong, Macau, Taiwan, Japan and South Korea.

Western populations include people who share Judeo-Christian heritage, specifically, those from countries or regions in Europe and North America, and other ones typically influenced by Judeo-Christian culture such as Australia and New Zealand (Chaves, 2002; Hahm, 2003; Ng, 2003). To a large extent, East Asians and Western populations are each treated as monolithic groups in cross-cultural studies of creativity. Theoretically, East Asians have a common cultural heritage of Confucianism (and Taosim, Buddhism), and Western populations have shared Judeo-Christian culture (Averill et al., 2001; Wonder & Blake, 1992). The former are often treated as collectivistic societies and the latter as individualistic ones (Hofstede & Bond, 1984; Triandis, 1995). From the viewpoint of Niu and Sternberg (2006), Western culture tends to be individualistic, and self-worth is based on defying the crowd and developing the self, thus their emphasis on creativity is originality. Alternatively, Eastern culture is collectivistic and one's sense of value is built upon embracing the crowd to find one's position in society, thus their view of creativity is not defying the crowd.

Empirically, there are numerous cross-cultural studies comparing East Asians and Western populations on creativity, of which most employed DT tests as creativity measures (Guo et al., 2021; Niu & Sternberg, 2002). In general, there seems more evidence supporting the higher creativity level of Western respondents. For example, Saeki and colleagues (2001) showed that Japanese college respondents performed significantly worse on DT tests than those from America. Deng et al. (2016) found that Mainland Chinese college students performed significantly lower on the Abbreviated Torrance Test for Adults (ATTA) than their American peers. In another instance, Jellen and Urban (1989) reported that children from America, Germany and England had higher total creativity scores than their counterparts from China after studying 11 countries applying the Test for Creative Thinking-Drawing Production. However, there is also opposing evidence. For example, Cheung et al. (2016) documented higher DT scores of Chinese children compared to their counterparts from France. Similarly, Ball and Torrance (1978) showed that participants from Eastern cultures (e.g., Japan, China) had higher performance on internal visualizations (one scoring dimension of TTCT figural, Activity 3) when compared to peers from Western cultures (e.g., the US, Norway, and Germany).

In addition, culture can exert an influence on the processes and evaluation criteria of creativity (Karwowski, 2016; Niu & Sternberg, 2006). The standard definition of creativity emphasizes both originality and appropriateness of ideas or products (Runco & Jaeger, 2012). These are consistent with the two processes involved in creativity—idea generation and idea evaluation (Chua, et al., 2015). The former focuses on the generation of many different and potentially new ideas, whereas the latter emphasizes the evaluation and selection of original and potentially appropriate responses. Numerous researchers have suggested that Westerners think highly of originality and their Eastern counterparts attach more importance to appropriateness or usefulness (Adair & Xiong, 2018; Morris & Leung, 2010; Nijstad et al., 2010; Xie & Paik, 2019). This divergence in priorities can potentially be explained by the different philosophical root of the two cultures (Niu & Sternberg, 2006). Western culture is individualistic and regards creativity as a kind of divine inspiration, thus thinking of creativity as something totally new. The things created are totally different from existing ones thus their emphasis is on originality. In contrast, Eastern culture is mainly based on Confucianism, which thinks highly of family, benevolence, education, and hierarchical relationships (Chen & Chung, 1994; Kim, 2009). The things created should follow moral goodness and traditions thus their focus is on appropriateness. Creativity here is viewed as change or improvement of existing ideas or products rather than the generation of completely novel ideas. It appears that evaluative thinking is more important in

Eastern cultures (Niu & Sternberg, 2006), and this may cause cross-cultural differences in evaluative skill.

As mentioned in previous sections, compared to studies on DT, those on evaluative thinking are generally lacking (Runco, 2020; Silvia, 2008), particularly in cross-cultural contexts. A most recent study (Ivancovsky et al., 2019) was considered the first to investigate the cross-cultural differences in evaluative thinking (between Japanese, Korean and Israelis), but its focus was on evaluative stringency rather than evaluative accuracy. Specifically, the researchers investigated individuals' tendency to adopt lenient or strict criteria in judgment, that is, the tendency to consider ideas as more or less original, appropriate, etc., but did not examine cross-cultural differences in the accuracy of these creativity judgments.

Although it is too early to draw firm conclusions about cross-cultural differences in creativity between East Asians and Western populations, many studies have suggested that cultural orientation, especially individualism vs. collectivism, explains the observed differences (Ng & Smith, 2004). The following sections review individualism vs. collectivism as well as other conceptualizations of cultural orientation in the existing literature.

Cultural Orientation: Individualism-collectivism

According to Hofstede (1980b), individualism refers to a loosely connected social system in which people are assumed to care for themselves and immediate families, whereas collectivism represents a tightly connected system in which in-groups (e.g., relatives, clan, organizations) and out-groups can be distinguished, and people are loyal to in-groups. Hofstede (1980a) published the Individualism Index Values (1^{st} edition) covering 50 countries as well as three regions. Among them, Western countries such as the US, Canada, and Australia were high on individualism, whereas East Asian countries or territories such as Taiwan and Korea ranked low, indicating dominant collectivistic values in those societies. Individualism vs. collectivism was thought to be the most important aspect of cultural variation (Triandis, 1995). A review of the literature indicated that after cultural differences in creativity were found, it is popular for researchers to use the logic of individualism vs. collectivism to explain the differences (Adair & Xiong, 2018; Erez & Nouri, 2010).

Specifically, individualism focuses on independence and self-initiative, which should be critical to novelty (Jones & Davis, 2000) whereas collectivism focuses on interdependence and conformity, which may hinder unique idea generation (Brew & Chen, 2007; Ng, 2001). Moreover, most researchers tend to regard respondents from one country as individualistic and the other as collectivistic and then use nationality as proxy to investigate the relationship between individualism vs. collectivism and creativity without directly measuring group differences in individualism vs. collectivism (Xie & Paik, 2019). The problem is that there are many other factors uncontrolled such as multicultural experiences, SES, and measurement errors (Cheung et al., 2016; Fee & Gray, 2012; Guo et al., 2021; Kim, 2016), and few empirical studies have systematically investigated whether cultural differences in creativity are indeed due to the individualism vs. collectivism difference. A notable exception is a study by Kim (2016), who directly investigated the role of individualism vs. collectivism (with a quantitative measure) in predicting divergent thinking among Korean college students. This study found no significant effect, possibly because the sample was culturally homogeneous, and thus, future studies should focus on groups from different cultures or subcultures to further investigate the issue.

It should be noted that the concept of individualism vs. collectivism has been further refined by incorporating the horizontal vs. vertical dimension of social relationships (Singelis et al., 1995; Triandis, 1995). The horizontal tendency assumes that everyone is equal and similar to others, and the vertical just the opposite. Horizontal individualism (HI) refers to that individuals want to be independent from groups, but everyone is equal to others in status, whereas in vertical individualism (VI) people would like to acquire status and be distinguished. Horizontal collectivism (HC) refers to that people have common goals and interdependent, but are similar to each other in status, whereas in vertical collectivism (VC) people are different from each other in status (e.g., ranks) despite sharing common goals and being dependent on one another. Studies to date have demonstrated that this four-dimensional individualism-collectivism taxonomy can meaningfully predict a number of creativity-related outcomes. For example, research showed that differences in these four dimensions mediated the relationship between cultures (Germany vs. Poland) and creativity mindset preference (growth vs. fixed), which are important to creativity-related outcomes (Tang et al., 2016). In another study (Yao et al., 2012) investigating Chinese employees, both HI and HC positively predicted idea generation, whereas VC positively affected idea implementation.

Cultural Orientation: Multiple Dimensions

Another influential conceptualization of cultural orientation concerns a group of four dimensions proposed by Hofstede (1980a) in the context of cross-cultural differences in work-related values. Besides individualism vs. collectivism, which was discussed above, the other three dimensions include power distance, uncertainty avoidance, and masculinity (vs. femininity). Research has shown that this last dimension is not associated with innovation (Rinne et al., 2012; Shane, 1993), so it will not be discussed further. A high level of power distance indicates that there is inequality in social systems and people are assumed to conform with authority, which may inhibit people from proposing novel ideas (Erez & Nouri, 2010; Morris & Leung, 2010). Uncertainty avoidance represents the extent to which people tend to feel threatened in unstructured situations (Hofstede, 1991). People from cultures with a high level of uncertainty avoidance feel more threatened by ambiguous situations, and thus, it is more likely that novelty will be avoided (Adair & Xiong, 2018).

The multiple dimensions of cultural orientation are theoretically sound in predicting creativity, but empirical studies investigating this issue are lacking. Among the small number of relevant studies, Adair and Xiong (2018) empirically investigate the mediating role of cultural orientation in the relationship between cultures and creativity. They found that uncertainty avoidance, rather than individualism-collectivism or power distance, mediated the relationship between cultures and creativity conceptions (i.e., novelty vs. usefulness). In another instance, Rinne and colleagues (2013) analyzed traits of different cultures through the perspective of the four dimensions and found that individualism vs. collectivism was the only dimension that could predict a nation's creativity.

Cultural Orientation: Confucianism and non-Confucianism

The third conceptualization on cultural orientation is Confucianism vs. non-Confucianism (Kim, 2009; Tan, 2016), which is less commonly researched than the previously mentioned dimensions of cultural orientation. As mentioned in the section regarding the definition of the two groups, East Asians are culturally different from Western populations because the former group of people are greatly influenced by Confucianism. From the perspective of Chen and Chung (1994), Confucianism can be summarized as an emphasis on education (e.g., rote learning, hard work), family system (e.g., obedience, filial piety), hierarchical social relationships (e.g., respecting elders, gender inequality), and benevolence (e.g., suppression of emotions, conformity). Kim (2009) investigated Confucianism (measured by Eastern-Western Perspective Scale) and creativity (by TTCT) on Korean educators, and results showed a negative association between the two. Specifically, elements including gender role expectations, gender inequality, unconditional obedience, and suppression of emotions may hinder creativity. Similarly, using the same measures, Kim et al. (2011) studied Korean and American respondents' level of Confucianism and creativity. They found that the former group were strikingly more Confucian than the latter. In addition, the American group were more adaptively creative and have more creative strengths than the Korean group, whereas Korean respondents were significantly more innovatively creative than American peers. Further, the Confucian elements including filial piety, gender role expectation, gender inequality, and suppression of expression were negatively associated with creativity.

In summary, the former two conceptualizations of cultural orientation are largely based on data-driven research such as what Hofstede (1980a) did by recruiting respondents from all over the world and the Individualism Index Value for each country. Different from the former two conceptualizations, the Confucianism vs. non-Confucianism is more reflective of the typical mainstream East Asian culture. It might be more applicable to daily life in various domains, but the concept itself can be too complex.

Based on the analysis above, there seems a general tendency that cultural orientation can predict creativity, but which dimension (s) can have an impact remains unclear. Besides the external factors mentioned in interpreting cross-cultural differences, such as cultural orientation, multicultural experiences, and SES (Cheung et al., 2016; Fee & Gray, 2012; Guo et al., 2021; Kim, 2016), researchers also proposed possible measurement issues underlying the assessment of creativity cross-culturally (Karwowski, 2016; Runco, 2004). Specifically, people from different cultures may have different conceptions of creativity (Cramond et al., 2020; Niu & Sternberg, 2002; Shao et al., 2019), and it is important to ensure any administered assessment measures the same underlying construct. There is increasing evidence showing that when culturally appropriate measures were used Easterners showed superior performance on creativity outcomes (Chen et al., 2004; Chua et al., 2015). Therefore, it is of critical importance to employ measures that are invariant across cultures in order to validly investigate cross-cultural differences in both DT and evaluative skill. Measurement invariance is reviewed in detail in the following section.

Measurement Invariance (MI)

Measurement invariance concerns the issue whether the indicators or items of a measure are understood in the same way across different conditions (Meade & Lautenschlager, 2004). These different conditions can include different time points, measurement methods, or populations. It is worth noting that in the Item Response Theory (IRT) framework, lack of MI refers to the concept of differential item functioning (DIF). In other words, if MI does not hold, it indicates existence of DIF, in which case item responses relate differently to latent constructs across groups (Bauer, 2017; Belzak & Bauer, 2020). Specifically, in terms of cross-cultural studies, MI deals with the issue whether the indicators or items of a measure mean the same thing among various cultural groups. For example, in the case of evaluative skill, it is critical to ensure MI is satisfied for measures of evaluative skill across different cultures. If the assumptions of MI are met, it would be valid for researchers to attribute higher observed scores of one group to their higher level of evaluative skill as a latent trait. Otherwise, the group differences in observed scores may just be a reflection of biased measurements (Wicherts et al., 2005).

In terms of invariance, researchers distinguished between measurement invariance and structural invariance (Byrne et al., 1989). The former concerns whether the indicators measure the latent traits in the same way across groups, whereas the latter concerns the latent traits being measured relate and are distributed the same across groups, that is, whether they have the same factor means, factor variances and factor covariances. MI established at a certain step makes it legitimate to test structural invariance (e.g., comparing latent factor means), which will be introduced later. For measurement invariance, there are four types or steps: configural invariance, weak invariance, strong invariance, and strict invariance (Wu et al., 2007). Based on existing MI literature (Meade & Lautenschlager, 2004; Said-Metwaly et al., 2020b), (a) configural invariance requires that the number of factors and the factor-indicator relationships are the same across different groups. It is the basic level that MI should hold. (b) With configural invariance established, the next level of MI is weak invariance, also called pattern or metric invariance. Weak invariance further assumes that factors loadings (or unstandardized pattern coefficients) should be invariant across different groups. In other words, the slopes obtained by regressing items on latent factors should be the same across groups. (c) Based on the satisfaction of weak invariance, the next level of MI, strong (or scalar) invariance, further assumes that the intercepts obtained by regressing items on latent factors should be the same across groups. In other words, individuals of one group would obtain the same observed score on an item as peers of another group with the same level of latent ability. (d) Last, with strong invariance satisfied, the highest level of MI, strict invariance further assumes equal error covariances and variances across different groups. Satisfaction of MI at this level means that indictors or items of a measure reach the same precision across different groups.

To determine the degree to which the items mean the same in different groups, testing of MI includes a series of analyses, which provides evidence for construct validity across groups (Cordon & Finney, 2008; Garcia-Barrera et al., 2011). In general, to test whether MI at each step is established, the free baseline (or bottom-up) approach was applied (Stark et al., 2006), that is, a group of nested models with different levels of cross-group equality constraints gradually added were fitted and compared by means of examining the model $\chi2$ change. It is worth noting that examining the model $\chi2$ change to determine MI may have some bias with a large sample size, leading to the rejection of MI. To address this problem, other criteria were suggested, for example, a -.01 in ΔCFI (Cheung & Rensvold, 2002; Xu & Barnes, 2011). Besides the four types above, researchers have also proposed the concept of partial invariance (Byrne et al., 1989). For example, if some but not all factor loadings are the same, partial weak invariance is considered met, and if some but not all intercepts are the same, partial strong invariance is satisfied. Based on Thompson and Green (2006), at least partial strong MI should be established for meaningful latent mean comparisons. In case partial strong MI cannot be achieved, theoretically and empirically driven item reduction (i.e., with some problem items removed) would be conducted to produce an invariant scale to conduct latent mean comparisons. This approach has also been used in other studies (e.g., Hansen et al., 2014; Sturm et al., 2017; Williams & Gotham, 2021). For more details on the procedure for MI tests at each step and descriptions on item dropping, see the section of Data Analysis Strategy under the Method chapter.

Thus far, research investigating MI issues in the field of creativity is rare, although a small number of studies have attempted to do so (Kim et al., 2006; Krumm et al., 2014; Krumm et al., 2016; Kuhn & Holling, 2009; Said-Metwaly et al., 2020b). Specifically, Kim and colleagues (2006) tested MI for the Torrance Tests of Creative Thinking (TTCT) figural scores and supported a two-factor latent model—adaptive and innovative, and MI was achieved to a greater degree across genders than across grades. Similarly, Krumm et al. (2016) tested the MI of TTCT-figural scores in Spanish-speaking children, and results supported a two-factor model, innovative and adaptive, and strong MI was supported across genders. In another instance, Kuhn and Holling (2009) tested MI for 12 DT tests (e.g., Uses, symbol combining) fluency scores and supported a three-factor model: verbal, figural and numerical; and partial strong MI was satisfied across age groups, genders and school forms. In addition, Said-Metwaly and colleagues (2020b) investigated TTCT-verbal scores in Egyptian college students and results supported a bifactor latent structure. In terms of MI, it suggested satisfaction of strict MI among groups of different academic majors, gender, and year of study. A recent study (Guo et al., 2021) that tested MI of two DT tests (Line Meanings and Real-world Problems) among American and Chinese college students was one of the first to investigate MI issues of DT measures in cross-cultural contexts. Results supported a two-factor model based on the two types of DT tests; weak invariance was supported for both fluency and originality indices; and partial strong MI was satisfied for Real-world Problems fluency only.

In sum, it seems that different factor models can be established based on different DT tests and DT indices, and that weak MI of DT tests can often be satisfied, although strong MI may be less common. Currently, studies testing MI in the creativity field are lacking (Karwowski, 2016), and more future studies should examine this issue. To investigate cross-group differences on a construct of interest, for example, cross-cultural differences on evaluative skill, traditionally researchers would compare observed means with ANOVA or t-test. However, to reach more statistically robust conclusions, it is suggested that latent mean comparisons be conducted to control for measurement bias (Barbot, 2019; Finch & French, 2015; Yuan & Bentler, 2006). Based on Thompson and Green (2006), at least partial strong MI should be satisfied for meaningful latent mean comparisons. In addition, if a study aims to investigate the relationships between different constructs across groups, weak MI needs to be satisfied (Chen & West, 2008; Rusticus & Hubley, 2006). For example, to examine whether the direction and strength of the relationship between individualism-collectivism and evaluative skill is comparable in American and Chinese groups, weak MI of the measures of both constructs needs to be satisfied.

CHAPTER 3

METHOD

This chapter first introduces the participants recruited for the present study, followed by the measures employed for each construct: divergent thinking tests, evaluative measures, the individualism-collectivism test and the uncertainty avoidance test. Next, it introduces the data analysis strategies used for the present study: the estimation method, the fit indices, tests of measurement invariance, and structural equation models analysis.

Participants

Data were collected in a public university in the central part of China, and a public university in the southeastern United States. In total, there were 345 college students (288 females, 83.5 %; mean age 20.25, *SD* = 2.658) in the Chinese group and 341 college students (243 females, 71.3 %; mean age 20.65, *SD* = 2.514) in the American group. The American sample included 61.9 % White, 7 % Latino, 1.2 % African American, 19.4% Asian, 6.5 % Native Hawaiian or Pacific Islander, and 4.1% multiracial and other individuals. Individuals from both groups were from a wide range of majors, including art, science, and social science. People from both cultural groups participated in an online survey containing the below-described measures in the following order: divergent thinking tests, the individualism-collectivism test, evaluative measures, the uncertainty avoidance test, and demographic measures.

Measures

Divergent Thinking Tests (see Appendices A-D)

Four kinds of DT tests were used in the study: Line Meanings, Uses, Instances, and Consequences. Among them, Line Meanings, Uses, and Instances were discussed by Wallach and Kogan (1965); the Consequences test was discussed by Guilford and colleagues (Christensen et al., 1953; Guilford, 1967, 1979). There were two subtests employed for each kind of test. For Line Meanings, participants were asked to imagine things what each drawing could be (a hook-like figure; a tie-like figure); for Uses, they were to think of uses for each daily object (a chair; a rope); for Instances, there were prompted to think of instances with a certain feature (square; move on wheels); for Consequences, they were required to list consequences of each unusual situation (if people lost the ability of reading and writing; if people no longer needed food in life). For each kind of test, participants were instructed to provide as many ideas as possible, and an example item and potential answers were presented beforehand to facilitate their understanding on how to respond. They approached the tasks like a game with unlimited time.

Objective DT indices (or outputs) were adopted to score the responses, including fluency and originality. With this method, fluency and originality scores were calculated for each respondent on each subtest. Fluency scores were calculated by counting the number of total relevant responses, and originality scores were obtained by counting the total number of the responses that were original (i.e., ideas given by < 10 % of the sample in each cultural group; Plucker et al., 2014).

Evaluative Measures (see Appendices E-H)

As mentioned in the section of literature review, to measure evaluative skill, there are two stages: In Stage One, ideas (usually from divergent thinking tests) are produced, and they are scored in terms of originality or creativity, which can then serve as the standard (or criterion, right answer) afterwards. In Stage Two, respondents are required to rate the ideas presented in terms of originality or creativity, and those ratings are compared to the objective criterion.

In the present study, Stage One was a bit different; ideas to be rated were collected from previous studies using DT tests of Line Meanings, Uses, Instances, and Consequences among American and Chinese groups (Guo & Guo, 2021; Guo et al., 2021; Silvia, 2008; Zhao, 2016; in studies employing Line Meanings and Uses, there were over 300 participants from each cultural group; in those employing Instances and Consequences, the sample size ranged from 52 to 242). They were scored in terms of originality (the percentage of people who proposed the idea out of the whole sample), which served as the criterion afterwards for calculating the evaluative accuracy index. In Stage Two, participants were asked to rate the ideas compiled in terms of originality (see detailed descriptions in the next paragraph). It should be noted that the types of evaluation tasks were the same with the DT tests used in the present study as mentioned above; however, the subtest content was different to reduce possible carryover or learning effect. Specifically, for idea evaluation tasks based on Line Meanings, there were another two figures (a curve and point figure; a wave-like figure); for Uses, another two daily objects (an umbrella; a book); for Instances, another two features (round; make a noise); for Consequences, another two situations (if people no longer needed to sleep; if people could live twice longer than now).

For idea compiling for evaluative measures, following Runco and colleagues' method (Charles & Runco, 2001; Runco & Vega, 1990), based on response statistical frequency in previous studies, two highly original ideas (i.e., given by less than 5% of the sample), two moderately original ideas (i.e., given by greater than 5% but less than 15% of the sample), and two common ideas (i.e., given by more than 15% of the sample) were selected from the idea pool for each of the eight subtests from each cultural group. It should be noted that the six ideas selected may vary depending on the cultural group, although some were the same. In total, there were 48 ideas for evaluation and within each subtest the six ideas were presented in random order. For all 48 ideas, participants were asked to rate the ideas on a 1-10 scale: How many people out of 10 do you think can give this idea? Similar task instructions have been used by previous researchers (Charles & Runco, 2001; Grohman et al., 2006).

Evaluative skill was operationalized as the rating accuracy. To obtain the accuracy index of originality ratings, the difference between participants' ratings and the criterion was obtained (Grohman et al, 2006). Specifically, the accuracy index was calculated by subtracting the originality criterion (the percentage of people who proposed the idea out of the whole sample in previous studies) from respondents' ratings (percentage value; for example, if a respondent gave the answer that 4 out 10 people can think of the idea, the value was 40%) on each presented idea to obtain an absolute value, which represented the discrepancy of judgement from the correct value. In this sense, higher values of discrepancy indicate lower levels of evaluative accuracy.

The Individualism-collectivism Test (see Appendix I)

The scale based on Triandis & Gelfland's (1998) work was used in the present study to assess horizontal and vertical collectivism and individualism. A 9-point scale (1 = disagree; 9 = agree) was used in the present study with 16 items in total covering four dimensions (four items for each): horizontal individualism (HI; e.g., "I often do my own thing."), vertical individualism (VI; e.g., "Winning is everything."), horizontal collectivism (HC; e.g., "If a coworker gets a prize, I would feel proud."), and vertical collectivism (VC; e.g., "It is my duty to take care of my family, even when I have to sacrifice what I want"). Cronbach's alpha was .65 (HI), .71 (VI), .72 (HC), .72 (VC) respectively in the study of Tang et al. (2016) with students from Poland and

Germany. Yao et al. (2012) also used that measure among Chinese employees with a 7-point scale. After confirmatory factor analysis (CFA), HI had 3 items (α = .64), VI had 2 items (α = .75), HC had 4 items (α = .76), and VC had 3 items (α = .70).

The Uncertainty Avoidance Test (see Appendix J)

Jung's (2002) measure was used, which was modified from the uncertainty avoidance scale by Hofstede's (1980a). A 7-point scale (1 = strongly disagree; 7 = strongly agree) was used in the present study including six items (e.g., "I feel stressful when I cannot predict consequences"). The reliability was good in a previous study (α = .82; Adair & Xiong, 2018).

Data Analysis Strategy

Estimation Method

All confirmatory factor analyses (CFA) analyses were conducted with the software Mplus, version 8.4 (Muthén & Muthén, 1998-2017). The default method of maximum likelihood estimation was used because the data in the present study were considered continuous. Specifically, for DT tests, the number of total ideas (i.e., fluency) or original ideas (i.e., originality) were counted; for evaluative measures, participants rated ideas on a scale from 1 to 10; they rated their agreement on each statement from 1 to 9 on the individualism-collectivism test, and from 1 to 7 on the uncertainty avoidance test. It is should be noted that there is debate on whether data from a Likert scale can be regarded as continuous (Harpe, 2015; Sullivan & Artino, 2013); however, researchers suggested that when the number of categories being selected exceeds or equals to five, the data can generally be regarded as continuous (Harpe, 2015). Thus, the statistical analyses conducted in the present study were acceptable, understanding that some findings may need to be interpreted with caution, especially for the individualism-collectivism test and the uncertainty avoidance test.

Fit Indices

In terms of overall fit, there are two types of fit indices: the absolute fit indices and the incremental fit indices (Hoyle & Panter, 1995; Hu & Bentler, 1999). The former indicates the degree to which the model implied covariances match the observed covariances. To serve this purpose, the chi-square test is employed, with nonsignificant statistic suggesting an adequate model (e.g., p >.05). It does not use any baseline model for comparison; instead, absolute fit indices are derived from the match between the implied covariances and the observed covariances based on maximum likelihood estimation. A p value larger than .05 was applied here to establish an acceptable model fit, although chi-squares may have bias with a large sample size, leading to model rejection. Besides chi squares, commonly used absolute indices also include others, such as the standardized root mean square residual (SRMR), which are simply transformed from chi-squares, with values less than .08 indicating an acceptable fit. The incremental fit indices indicate the degree to which the current model being tested is better than a baseline model (i.e., no relations among variables), with higher values implying greater improvement of the current model compared to a baseline. The following index is recommended: the comparative fit index (CFI), with values above .90 suggesting an adequate model fit (Browne & Cudeck, 1993). Last, another fit index is suggested to indicate the lack of fit between the model and the sample data: the root mean squared error of approximation (RMSEA), with values less than .08 considered acceptable (Browne & Cudeck, 1993). It is worth noting that these cutoff values for CFI and RMSEA are relatively liberal, understanding that there is debate on this and other researchers suggested more strict criteria, for example, CFI > .95 and RMSEA < .06 (Hu & Bentler, 1999). In sum, the present study used the following indices with liberal criteria to assess the model fit: chi-square statistic, CFI, RMSEA, and SRMR.

Tests of Measurement Invariance: Multiple-Group CFA Analysis

In general, to test whether MI at each step is established, the free baseline (or bottom-up) approach was applied (Stark et al., 2006), that is, a group of nested models with different levels of cross-group equality constraints gradually added were fitted and compared by means of examining the model $\chi 2$ change. First, configural invariance was tested by freeing all parameters across the two groups, which was the unconstrained baseline model. For each indicator-factor relationship, if the factor loading estimated across the two groups shows the same sign (e.g., both positive, or both negative, or both zero), in addition to a good model fit, it suggests the satisfaction of configural invariance. Second, when configural MI was met, weak and strong MI were tested step by step. Specifically, based on the unconstrained model, different levels of cross-group equality constraints were gradually added, including factor loadings and intercepts. If at a certain step, the more constrained model turns out to be significantly worse, then the current model is not acceptable and will be discarded, indicating the unsatisfaction of the hypothesized level of invariance. It should be cautioned that examining the model $\chi 2$ change to determine MI may have bias with a large sample size, leading to the rejection of MI. To solve this problem, other criteria have been suggested, for example, a -.01 in ΔCFI (Cheung & Rensvold, 2002; Xu & Barnes, 2011). In other words, when compared with the less restricted model, if the reduced CFI value is less than or equals .01, the more constrained model will be retained since it is more parsimonious. Third, to effectively conduct latent mean comparisons, if strong MI was not met, partial strong MI would be tested as an option (Thompson & Green, 2006). Since the ultimate goal of the present study was to investigate differences on constructs of interest across groups via latent mean comparisons, strong or partial strong MI needed to be met (Thompson & Green, 2006). It also aimed to investigate the relationships between constructs across groups, thus in this case, weak MI needs to be satisfied (Chen & West, 2008; Rusticus & Hubley, 2006). In light of this, when MI could not be met, theoretically and empirically driven item reduction (i.e., with some problem items removed) would be conducted to reach the desired degree of MI (weak, strong or partial strong invariance) to produce a reduced but invariant scale. This approach has also been used in other studies (e.g., Hansen et al., 2014; Sturm et al., 2017; Williams & Gotham, 2021). In the process of confirming the latent factor structure, when there was a poor model fit, modification indices were examined first. Specifically, in Mplus, modification indices > 10 indicate parameters that may strongly affect the model fit. However, suggestions by modification indices are data driven, and when there is conflict between data and theories, the latter should prevail (Frazier & Youngstrom, 2007; Garcia-Barrera et al., 2011). In addition, Exploratory Factor Analyses (EFA) was conducted if needed. Based on these results and theories, problem items were detected and discarded. The same procedure of detecting problem items was followed in the MI tests after the latent structure was determined.

Structural Equation Models Analysis

When weak MI of the measures for each construct was established, that is, the measurement models were legitimate to investigate relationships between constructs across groups, a general SEM model was fitted for each cultural group. The present study aimed to investigate the relationships between individualism-collectivism, uncertainty avoidance, divergent thinking, and evaluative skill in American and Chinese individuals. The theoretical relationships in the structural models to be tested were as follows: individualism-collectivism and uncertainty avoidance served as the predictor variables; divergent thinking and evaluative skill served as the outcome variables. The fit indices used for assessing the model fit were the same as in CFA models mentioned above, including chi-square statistic, CFI, RMSEA, and

SRMR. Two other absolute fit indices were also added for estimation, including AIC (Akaike's Information Criterion) and BIC (Bayesian Information Criterion).

CHAPTER 4

RESULTS

This chapter reports the analysis results for all research questions. Table 1 shows exactly which steps of analyses corresponded to each research question. In the following formal analysis, for measures for each of the constructs (divergent thinking, evaluative skill, individualism-collectivism, and uncertainty avoidance), the latent factor structure was explored and confirmed first, and different steps of MI (configural, weak, and strong) were tested, followed by latent mean comparisons to examine the differences in DT as well as evaluative skill across the two cultural groups. Last, general SEM models were fitted to investigate the relationships between all variables: individualism-collectivism, uncertainty avoidance, DT, and evaluative skill.

Table 1
Research Questions and Steps of Analyses

Question number	Research question	Step of analysis	References
	Data coding before formal analysis	Scoring/Coding for divergent thinking tests: Each participant will get a fluency score and an originality score on each of the eight subtests. *Fluency*—the number of total ideas produced; *originality*—the number of original ideas, that is, ideas proposed by <=10% of the sample.	Plucker et al., 2014
	Data coding before formal analysis	Coding for evaluative measures: The evaluative score was calculated by subtracting the originality criterion from respondents' ratings on each presented idea to	Grohman et al., 2006

		obtain an absolute value, which represented the discrepancy of judgement from the correct value.	
	Data coding before formal analysis	For individualism-collectivism test, and uncertainty avoidance test, participants' ratings on the Likert scale were directly used for analysis.	
Questions 1–4 were applied to measures for each of the constructs: divergent thinking, evaluative skill, individualism-collectivism, and uncertainty avoidance: For American and Chinese groups,			
1	What is the latent structure of the measures?	CFA: Confirm the number of factors in each group and in combined data	From Questions 1–7: Meade & Lautenschlager, 2004; Rusticus & Hubley, 2006; Said-Metwaly et al., 2020b; Wu et al., 2007
2	Do the measures have the same factors and relationship between indicators and factors across the two cultural groups?	Multigroup CFA: Testing configural measurement invariance (MI)	
3	For the measures, are factors linked with corresponding indicators to the same extent (strength) across the two cultural groups?	Multigroup CFA: Testing weak MI	
4	For the measures, do American respondents obtain the same observed score on an indicator with Chinese peers who have the same ability on a respective latent factor?	Multigroup CFA: Testing strong MI	
5	Are there differences between the two cultural groups in divergent thinking as latent factor (s)?	Multigroup CFA: Latent mean comparisons	
6	Are there differences between the two cultural groups in evaluative skill as latent factor (s)?	Multigroup CFA: Latent mean comparisons	
7	Is cultural orientation (individualism vs.	For each cultural group, SEM was used to model	

	collectivism, uncertainty avoidance) associated with divergent thinking and evaluative skill?	the relationships between all variables	

Divergent Thinking Tests Fluency Scores

Descriptives and Correlations between Variables

Table 2 displays means, standard deviations, correlations between all eight subtests of divergent thinking fluency scores (correlations in the American sample were below the diagonal; correlations in the Chinese sample were above the diagonal). Results showed positive correlations between each other. It should be noted that for each of the eight subtests, there was a high intercorrelation between the score of originality (another output of DT, see the measures section for details) and fluency (i.e., .704 ~ .917 in the American group and .644 ~ .887 in the Chinese group), which was consistent with the literature (Guo et al., 2018; Plucker et al., 2014). In light of this, DT originality scores were analyzed in a separate set. This approach has also been used in other studies (e.g., Guo et al., 2021).

Table 2

Descriptives and Correlations between Fluency Scores

	US		CN									
	M	SD	M	SD	1	2	3	4	5	6	7	8
LM1_flu	3.92	2.04	3.15	1.65	1	.721**	.487**	.464**	.487**	.316**	.384**	.394**
LM2_flu	3.70	1.93	2.90	1.51	.669**	1	.507**	.501**	.516**	.410**	.388**	.404**
Use1_flu	5.01	2.00	4.07	1.69	.527**	.450**	1	.594**	.524**	.511**	.406**	.463**
Use2_flu	5.03	2.64	4.31	2.17	.460**	.494**	.659**	1	.598**	.580**	.452**	.481**
Ins1_flu	7.39	4.10	6.69	3.88	.439**	461**	.564**	.534**	1	.594**	.426**	.452**
Ins2_flu	8.86	4.67	5.17	2.70	.402**	.472**	.522**	.589**	.613**	1	.535**	.538**
Co1_flu	4.60	2.60	3.26	1.79	.408**	.405**	.454**	.520**	.529**	.569**	1	.702**
Co2_flu	4.42	2.51	3.13	1.79	.404**	.417**	.382**	.482**	.444**	.572**	.698**	1

Note. American sample (US) n = 341; Chinese sample (CN) n = 345; correlations in the American sample were below the diagonal; correlations in the Chinese sample were above the diagonal; LM1_flu, LM2_flu, Use1_flu, Use2_flu, Ins1_flu, Ins2_flu, Co1_flu, Co2_flu refer to fluency scores on the first and second item on DT tests of Line Meanings, Uses, Instances, and Consequences respectively.

**$p < .01$.

Initial Reliability Analyses

The internal consistency coefficients were .801, .776, .756, .822 for Line Meanings, Uses, Instances, and Consequences respectively for the American group, and .836, .730, .715, .825 for the Chinese group, which were considered moderate to high reliability. In terms of the alpha coefficient over all eight subtests, the values were .870 for the American group and .863 for the China group, which indicated a potential multidimensionality of the measures for divergent thinking fluency.

Confirmatory Factor Analysis: Fluency Scores

To answer the first question on the latent structure of the measures, CFA models with different number of factors were tested. Figure 1 displays the theoretical models to be tested for fluency scores. The one-factor model assumed the eight subtests measured a universal concept or construct of DT, whereas the four-factor model assumed distinctiveness of each kind of DT tests, which was supported by the literature (Runco et al., 2016). Table 3 shows the chi square statistic and fit indices for the two potential factor models when factor loadings were freely estimated in each cultural group as in the configural MI model. There was difference in chi square but not tested since they were not nested models. In terms of model fit, all indices suggested that the one-factor model showed a poor fit (e.g., CFI = .842), whereas the four-factor model fit the data well (e.g., CFI = .981). The results supported that the eight subtests selected provide information about four components of DT. Thus, the following steps of testing MI were based on the four-factor model.

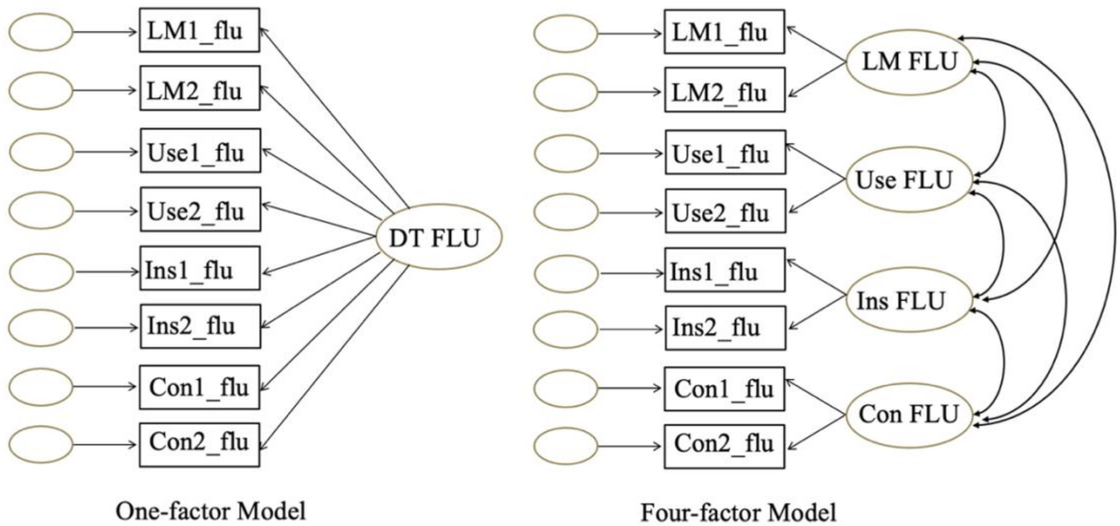


Figure 1. *Theoretical Models to be Tested: Fluency Scores*

Note. LM1_flu, LM2_flu, Use1_flu, Use2_flu, Ins1_flu, Ins2_flu, Con1_flu, Con2_flu refer to fluency scores on the first and second item on DT tests of Line Meanings, Uses, Instances, and Consequences respectively. DT FLU = DT fluency; LM FLU = Line Meanings fluency; Use FLU = Uses fluency; Ins FLU = Instances fluency; Con FLU = Consequences fluency.

Table 3

Results of Model Comparison of One- and Four-Factor Models: Fluency Scores

Model	No. of factors	χ^2	*df*	CFI	TLI	RMSEA	SRMR	$\Delta\chi^2$
Configural	One	474.165	40	.842	.778	.178	.066	
Configural	Four	79.478	28	.981	.962	.073	.026	393.687

Note. df = degree of freedom; CFI = Comparative fit index; TLI = Tucker Lewis index; RMSEA = Root mean square error of approximation; SRMR = Standardized root mean square residual.

Testing of Measurement Invariance

Parameter estimates based on the four-factor configural MI model are displayed in Table 4, and the model fit results for each step of MI tests are shown in Table 5. For the configural MI model, with the factor loading on the first item set at 1 while those on the second item freely estimated, results showed positive factor loadings estimated for each indicator-factor relationship in each cultural group. These estimates were significant; combined with the fit indices information indicating a good model fit (e.g., CFI = .981), these results together suggested the *configural invariance was supported* in the present analysis. In addition, the difference in CFI between the configural MI model and weak MI model was -.009, satisfiying the ΔCFI <= .01 criterion (Cheung & Rensvold, 2002), *supporting the assumption of weak invariance*. However, the difference in CFI between the weak MI model and strong MI model was -.037, exceeding the ΔCFI <= .01 criterion, suggesting *the violation of strong invariance*. Informed by modification indices based on the strong MI model, two subtests of Instances were potential problem items, followed by two subtests of Uses. With the two intercepts of Instances across the groups freely estimated (CFI = .943), partial strong MI was tested but not supported. Several other kinds of partial MI models were tested as well but they were not supported either.

Table 4

Configural MI Model Parameter Estimates on Fluency Scores across Two Groups: Four Factors

	LM FLU by		Use FLU by		Ins FLU by		Con FLU by	
	LM1_flu	LM2_flu	Use1_flu	Use2_flu	Ins1_flu	Ins2_flu	Con1_flu	Con2_flu
American	@1	0.963***	@1	1.381***	@1	1.214***	@1	0.907***
Chinese	@1	0.986***	@1	1.391***	@1	0.699***	@1	1.048***

Note. LM FLU = Line Meanings fluency; Use FLU = Uses fluency; Ins FLU = Instances fluency; Con FLU = Consequences fluency.

***$p < .001$.

Table 5

Fit Indices for Model Comparisons on Fluency Scores across Two Groups: Four Factors

Model	χ^2	*df*	CFI	RMSEA	SRMR	Compare with	$\Delta\chi^2$	Δdf	ΔCFI
Configural	79.478	28	.981	.073	.026				
Weak	109.070	32	.972	.084	.049	configural	29.592***	4	-.009
Strong	213.849	36	.935	.120	.071	weak	104.779***	4	-.037

Note. *df* = degree of freedom; CFI = Comparative fit index; RMSEA = Root mean square error of approximation; SRMR = Standardized root mean square residual.

***$p < .001$.

Next, item reduction was conducted to produce an invariant scale (e.g., Sturm et al., 2017; Williams & Gotham, 2021). Informed by the initial reliability analyses, the modification indices analyses above, and the nature of the eight subtests (i.e., every two subtests were based on a single type of DT tests), two subtests of Instances or two subtests of Uses were potential problem items to be removed to achieve strong MI. Specifically, the two items on the Instances test were dropped and the group of nested models were fitted again. To confirm whether the three-factor model remained the better model, the one-factor and three-factor models (see Figure 3) were tested again, and results are shown in Table 6 when factor loadings were freely estimated in each cultural group as in the configural MI model. Again, results suggested a significantly better fit of the three-factor model (e.g., CFI = .995) than the one-factor model (e.g., CFI = .806). The results further supported that the six subtests selected provide information about three components of DT. Thus, the following steps of testing MI were based on the three-factor model.

Parameter estimates based on the configural MI model are displayed in Table 7, and the model fit results for each step of MI tests are shown in Table 8. For the configural MI model, with the factor loading on the first item set at 1 while those on the second item freely estimated, results showed positive factor loadings estimated for each indicator-factor relationship in each cultural group. These estimates were significant; combined with the fit indices information indicating a good model fit (e.g., CFI = .995), these results together suggested the *configural invariance was supported* in the present analysis. In addition, the difference in CFI between the configural MI model and weak MI model was 0, meeting the ΔCFI <= .01 criterion (Cheung & Rensvold, 2002), *supporting the assumption of weak invariance*. Moreover, the difference in CFI between the weak MI model and strong MI model was -.005, additionally meeting the ΔCFI <= .01 criterion and *suggesting the satisfaction of strong invariance*.

Based on this strong MI model input, with the latent means in the American group set at zero while those in the Chinese group freely estimated, the latent mean comparison showed higher fluency of the American sample on all three types of DT tests: Line Meanings, Uses, and Consequences (*p* <.001).

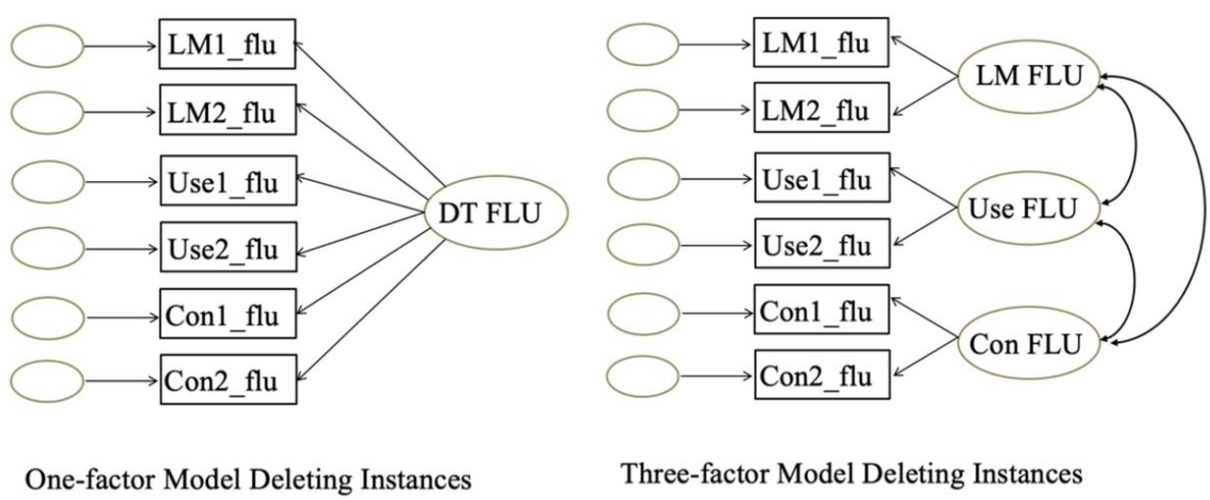


Figure 2. *Adjusted Theoretical Models to be Tested: Fluency Scores*

Note. The figure on the left refers to the one-factor model after omitting the two subtests of Instances, and that on the right represents the three-factor model after removing the two subtests of Instances. LM1_flu, LM2_flu, Use1_flu, Use2_flu, Con1_flu, Con2_flu refer to fluency scores on the first and second item on DT tests of Line

Meanings, Uses, and Consequences respectively. DT FLU = DT fluency; LM FLU = Line Meanings fluency; Use FLU = Uses fluency; Con FLU = Consequences fluency.

Table 6

Results of Model Comparison of One- and Three-Factor Models: Fluency Scores

Model	No. of factors	χ^2	*df*	CFI	TLI	RMSEA	SRMR	$\Delta\chi^2$
Configural	One	372.010	18	.806	.677	.239	.074	
Configural	Three	21.579	12	.995	.987	.048	.015	350.431

Note. *df* = degree of freedom; CFI = Comparative fit index; TLI = Tucker Lewis index; RMSEA = Root mean square error of approximation; SRMR = Standardized root mean square residual.

Table 7

Configural MI Model Parameter Estimates on Fluency Scores across Two Groups: Three Factors

	LM FLU by		Use FLU by		Con FLU by	
	LM1_flu	LM2_flu	Use1_flu	Use2_flu	Con1_flu	Con2_flu
American	@1	0.934***	@1	1.408***	@1	0.900***
Chinese	@1	0.964***	@1	1.305***	@1	1.080***

Note. LM FLU = Line Meanings fluency; Use FLU = Uses fluency; Con FLU = Consequences fluency.

***$p < .001$.

Table 8

Fit Indices for Model Comparisons on Fluency Scores across Two Groups: Three Factors

Model	χ^2	*df*	CFI	RMSEA	SRMR	Compare with	$\Delta\chi^2$	Δdf	ΔCFI
Configural	21.579	12	.995	.048	.015				
Weak	24.859	15	.995	.044	.020	configural	3.28	3	0
Strong	36.009	18	.990	.054	.029	weak	11.15*	3	-.005

Note. *df* = degree of freedom; CFI = Comparative fit index; RMSEA = Root mean square error of approximation; SRMR = Standardized root mean square residual.

*$p < .05$.

Divergent Thinking Tests Originality Scores

Descriptives and Correlations between Variables

Table 9 displays means, standard deviations, correlations between all eight subtests of divergent thinking originality scores (correlations in the American sample were below the diagonal; correlations in the Chinese sample were above the diagonal). Results showed positive correlations between all subtests.

Table 9

Descriptives and Correlations between Originality Scores

	US		CN									
	M	SD	M	SD	1	2	3	4	5	6	7	8
LM1_ori	1.86	1.61	1.58	1.41	1	.469**	.195**	.325**	.306**	.255**	.246**	.189**
LM2_ori	1.53	1.54	1.32	1.20	.536**	1	.342**	.393**	.301**	.376**	.288**	.268**
Use1_ori	1.99	1.64	1.07	1.10	.350**	.290**	1	.303**	.258**	.360**	.214**	.257**
Use2_ori	1.98	1.79	1.95	1.70	.287**	.330**	.415**	1	.373**	.410**	.298**	.324**
Ins1_ori	3.67	2.82	2.25	2.10	.275**	.302**	.424**	.425**	1	.378**	.349**	.355**
Ins2_ori	1.66	1.78	1.17	1.42	.242**	.291**	.339**	.317**	.420**	1	.485**	.417**
Co1_ori	3.42	2.44	2.37	1.66	.279**	.264**	.335**	.370**	.493**	.378**	1	.522**
Co2_ori	2.69	2.03	2.15	1.49	.288**	.202**	.279**	.326**	.346**	.398**	.597**	1

Note. American sample (US) n = 341; Chinese sample (CN) n = 345; correlations in the American sample were below the diagonal; correlations in the Chinese sample were above the diagonal; LM1_ori, LM2_ori, Use1_ori, Use2_ori, Ins1_ori, Ins2_ori, Co1_ori, Co2_ori refer to originality scores on the first and second item on DT tests of Line Meanings, Uses, Instances, and Consequences respectively.

$^{**}p < .01$.

Initial Reliability Analyses

The internal consistency coefficients were .697, .585, .549, .740 for Line Meanings, Uses, Instances, and Consequences respectively for the American group, and .633, .433, .519, .683 for the Chinese group, which were considered moderate reliability except for the relatively lower reliability of Uses. In terms of the overall alpha coefficient over all eight subtests, the values were .806 for the American group, and .791 for the Chinese group, which indicated potential multidimensionality of the measures for divergent thinking originality. It should be noted that the reliability analyses here are more of descriptives. The number of items for each subscale is only two, which may contribute to the relatively small alpha values of some measures. In other words, some values might be misleading. However, items with relatively smaller alpha values compared to others would be regarded as potential problem items that may influence the measurement invariance across groups.

Confirmatory Factor Analysis: Originality Scores

To answer the first question on the latent structure of the measures, CFA models with different number of factors were tested. Figure 3 displays the theoretical models to be tested for originality scores. The one-factor model assumed the eight subtests measured a universal concept or construct of DT, whereas the four-factor model assumed distinctiveness of each kind of DT tests, which was supported by the literature (Runco et al., 2016). Table 10 shows the chi square statistic and fit indices for the two potential factor models when factor loadings were freely estimated in each cultural group as in the configural MI model. There was difference in chi square but not tested since they were not nested models. In terms of model fit, all indices suggested that the one-factor model showed a poor fit (e.g., CFI = .867), whereas the four-factor model fit the data well (e.g., CFI = .993). The results supported that the eight subtests selected provide information about four components of DT. Thus, the following steps of testing MI were based on the four-factor model.

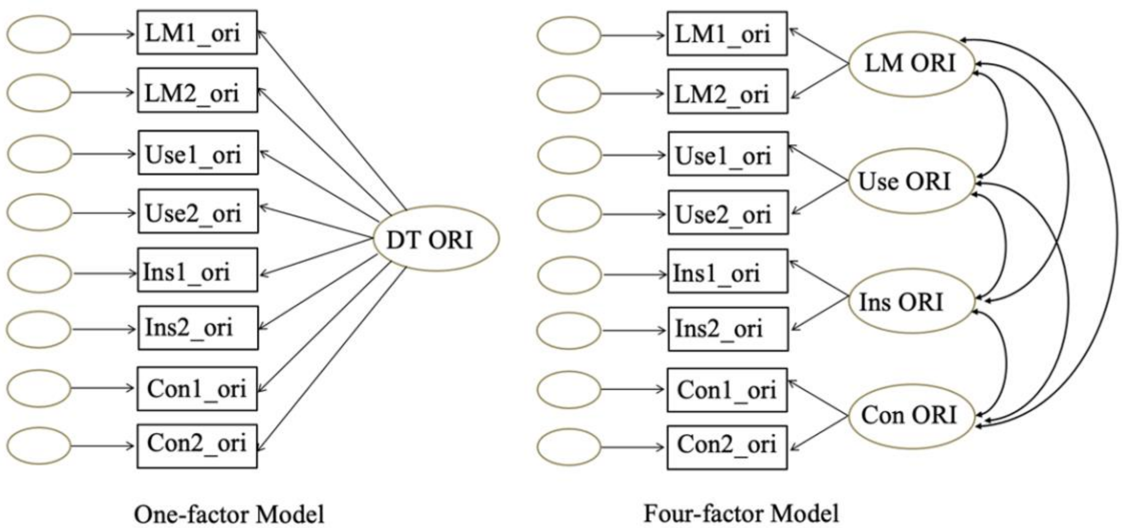


Figure 3. *Theoretical Models to be Tested: Originality Scores*

Note. LM1_ori, LM2_ori, Use1_ori, Use2_ori, Ins1_ori, Ins2_ori, Con1_ori, Con2_ori refer to originality scores on the first and second item on DT tests of Line Meanings, Uses, Instances, and Consequences respectively. DT ORI = DT originality; LM ORI = Line Meanings originality; Use ORI = Uses originality; Ins ORI = Instances originality; Con ORI = Consequences originality.

Table 10

Results of Model Comparison of One- and Four-Factor Models: Originality Scores

Model	No. of factors	χ^2	*df*	CFI	TLI	RMSEA	SRMR	$\Delta\chi^2$
Configural	One	221.059	40	.867	.814	.115	.058	
Configural	Four	37.011	28	.993	.987	.031	.020	184.048

Note. df = degree of freedom; CFI = Comparative fit index; TLI = Tucker Lewis index; RMSEA = Root mean square error of approximation; SRMR = Standardized root mean square residual.

Testing of Measurement Invariance

Parameter estimates based on the four-factor configural MI model are displayed in Table 11, and the model fit results for each step of MI tests are shown in Table 12. For the configural MI model, with the factor loading on the first item set at 1 while those on the second item freely estimated, results showed positive factor loadings estimated for each indicator-factor relationship in each cultural group. These estimates were significant; combined with the fit indices information indicating a good model fit (e.g., CFI = .993), these results together suggested the *configural invariance was supported* in the present analysis. In addition, the difference in CFI between the configural MI model and weak MI model was -.012, exceeding the ΔCFI <= .01 criterion (Cheung & Rensvold, 2002) and indicating *the violation of weak invariance*. Informed by modification indices based on the weak MI model, two subtests of Uses were potential problem items. That is, with the two factor loadings of Uses across the groups freely estimated, there could be a significant increase in the model fit. However, to conduct latent mean comparisons in the later stage, strong or partial strong invariance needed to be satisfied, and thus, modifications were made to the scale to reach this level of MI.

Table 11

Configural MI Model Parameter Estimates on Originality Scores across Two Groups: Four Factors

	LM ORI by		Use ORI by		Ins ORI by		Con ORI by	
	LM1_ori	LM2_ori	Use1_ori	Use2_ori	Ins1_ori	Ins2_ori	Con1_ori	Con2_ori
American	@1	0.951	@1	1.097	@1	0.523	@1	0.695
Chinse	@1	1.119	@1	1.932	@1	0.827	@1	0.837

Note. LM ORI = Line Meanings originality; Use ORI = Uses originality; Ins ORI = Instances originality; Con ORI = Consequences originality.

***$p < .001$.

Table 12

Fit Indices for Model Comparisons on Originality Scores across Two Groups: Four Factors

Model	χ^2	*df*	CFI	RMSEA	SRMR	Compare with	$\Delta\chi^2$	Δdf	ΔCFI
Configural	37.011	28	.993	.031	.020				
Weak	58.352	32	.981	.049	.040	configural	21.341***	4	-.012
Strong	126.519	36	.934	.086	.054	weak	68.167***	4	-.047

Note. *df* = degree of freedom; CFI = Comparative fit index; RMSEA = Root mean square error of approximation; SRMR = Standardized root mean square residual.

***p < .001.

Next, item reduction was conducted to produce an invariant scale (e.g., Sturm et al., 2017; Williams & Gotham, 2021). Informed by the initial reliability analyses, the modification indices analyses above, and the nature of the eight subtests (i.e., every two subtests were based on a single type of DT tests), two subtests of Instances or two subtests of Uses were potential problem items to be removed to achieve weak MI. Specifically, the two items on the Instances test were dropped and the group of nested models were fitted again. To confirm whether the three-factor model remained the better model, the one-factor and three-factor models (see Figure 4) were tested again, and results are shown in Table 13 when factor loadings were freely estimated in each cultural group as in the configural MI model. Again, results suggested a significantly better fit of the three-factor model (e.g., CFI = .998) than the one-factor model (e.g., CFI = .811). The results further supported that the six subtests selected provide information about three components of DT. Thus, the following steps of testing MI were based on the three-factor model.

Parameter estimates based on the configural MI model are displayed in Table 14, and the model fit results for each step of MI tests are shown in Table 15. For the configural MI model, with the factor loading on the first item set at 1 while those on the second item freely estimated, results showed positive factor loadings estimated for each indicator-factor relationship in each cultural group. These estimates were significant; combined with the fit indices information indicating a good model fit (e.g., CFI = .998), these results together suggested the *configural invariance was supported* in the present analysis. In addition, the difference in CFI between the configural MI model and weak MI model was -.008, meeting the ΔCFI <= .01 criterion (Cheung

& Rensvold, 2002) and supporting *the assumption of weak invariance*. However, the difference in CFI between the weak MI model and strong MI model was -.067, exceeding the ΔCFI <= .01 criterion and suggesting *the violation of strong invariance*. Informed by modification indices based on the strong MI model, two subtests of Uses were potential problem items. With the two intercepts of Uses across the groups freely estimated (CFI = .944), partial strong MI was tested but not supported. Several other kinds of partial MI models were tested as well but they were not supported either.

Further, informed by the initial reliability analyses, the modification analyses above, and the nature of the eight subtests (i.e., every two subtests were based on a single type of DT tests), two subtests of Uses were potential problem items to be further removed to achieve strong MI. Specifically, another model was tested by removing the two subtests of both Uses and Instances, and strong invariance was supported in this case. Parameter estimates based on the configural MI model are displayed in Table 16, and model fit indices for different levels of MI models are shown in Table 17. The difference in CFI between the configural MI model and weak MI model was .001, and the difference in CFI between the weak MI model and strong MI model was -.002, both meeting the ΔCFI <= .01 criterion (Cheung & Rensvold, 2002) and *supporting the assumption of weak and strong invariance*.

Based on the strong MI model input, with the latent means in the American group set at zero while those in the Chinese group freely estimated, the latent mean comparison showed higher originality of the American sample on two types of DT tests: Line Meanings and Consequences ($p < .05$; $p < .001$ respectively).

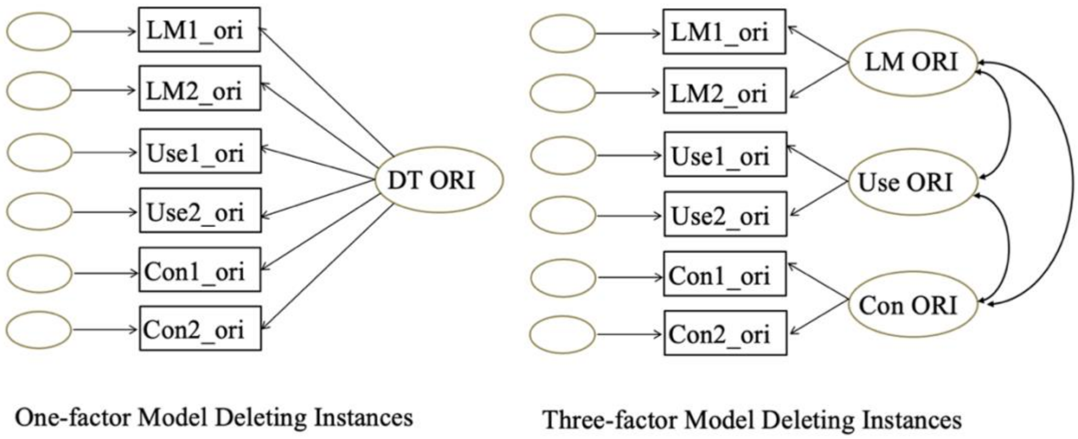


Figure 4. *Adjusted Theoretical Models to be Tested: Originality Scores*

Note. The figure on the left refers to the one-factor model after omitting the two subtests of Instances, and that on the right represents the three-factor model after removing the two subtests of Instances. LM1_ori, LM2_ori, Use1_ori, Use2_ori, Con1_ori, Con2_ori refer to originality scores on the first and second item on DT tests of Line Meanings, Uses, and Consequences respectively. DT ORI = DT originality; LM ORI = Line Meanings originality; Use ORI = Uses originality; Con ORI = Consequences originality.

Table 13

Results of Model Comparison of One- and Three-Factor Models: Originality Scores

Model	No. of factors	χ^2	*df*	CFI	TLI	RMSEA	SRMR	$\Delta\chi^2$
Configural	One	179.826	18	.811	.684	.162	.067	
Configural	Three	13.312	12	.998	.996	.018	.016	166.514

Note. df = degree of freedom; CFI = Comparative fit index; TLI = Tucker Lewis index; RMSEA = Root mean square error of approximation; SRMR = Standardized root mean square residual.

Table 14

Configural MI Model Parameter Estimates on Originality Scores across Two Groups: Three Factors

	LM ORI by		Use ORI by		Con ORI by	
	LM1_ori	LM2_ori	Use1_ori	Use2_ori	Con1_ori	Con2_ori
American	@1	0.903***	@1	1.138***	@1	0.728***
Chinese	@1	1.140***	@1	1.967***	@1	0.932***

Note. LM ORI = Line Meanings originality; Use ORI = Uses originality; Con ORI = Consequences originality.

***$p < .001$.

Table 15

Fit Indices for Model Comparisons on Originality Scores across Two Groups: Three Factors

Model	χ^2	*df*	CFI	RMSEA	SRMR	Compare with	$\Delta\chi^2$	Δdf	ΔCFI
Configural	13.312	12	.998	.018	.016				
Weak	23.913	15	.990	.042	.033	configural	10.601*	3	-.008
Strong	83.481	18	.923	.103	.058	weak	59.568***	3	-.067

Note. *df* = degree of freedom; CFI = Comparative fit index; RMSEA = Root mean square error of approximation; SRMR = Standardized root mean square residual.

*$p < .05$; ***$p < .001$.

Table 16

Configural MI Model Parameter Estimates on Originality Scores across Two Groups: Two Factors

	LM ORI by		Con ORI by	
	LM1_ori	LM2_ori	Con1_ori	Con2_ori
American	@1	0.802***	@1	0.783***
Chinese	@1	1.067***	@1	0.794***

Note. LM ORI = Line Meanings originality; Con ORI = Consequences originality.

***$p < .001$.

Table 17

Fit Indices for Model Comparisons on Originality Scores across Two Groups: Two Factors

Model	χ^2	*df*	CFI	RMSEA	SRMR	Compare with	$\Delta\chi^2$	Δdf	ΔCFI
Configural	3.221	2	.998	.042	.009				
Weak	4.350	4	.999	.016	.014	configural	1.129	2	.001
Strong	7.779	6	.997	.029	.019	weak	3.429	2	-.002

Note. *df* = degree of freedom; CFI = Comparative fit index; RMSEA = Root mean square error of approximation; SRMR = Standardized root mean square residual.

Evaluative Measures

Descriptives and Correlations Between Variables

For idea evaluation based on each of the eight DT subtests, the average evaluative scores were obtained over the six ideas evaluated. Therefore, for each respondent there were eight evaluative scores that were analyzed further. Table 18 displays means, standard deviations, correlations between variables (correlations in the American sample were below the diagonal; correlations in the Chinese sample were above the diagonal). The eight scores showed positive correlations between each other.

Table 18

Descriptives and Correlations between Evaluative Accuracy Scores

	US		CN									
	M	SD	M	SD	1	2	3	4	5	6	7	8
LME1_M	.36	.13	.37	.15	1	.588**	.453**	.410**	.505**	.333**	.349**	.284**
LME2_M	.55	.15	.54	.14	.551**	1	.508**	.450**	.471**	.308**	.338**	.338**
UseE1_M	.33	.15	.40	.14	.523**	.582**	1	.644**	.559**	.350**	.386**	.368**
UseE2_M	.41	.16	.45	.17	.405**	.473**	.636**	1	.550**	.397**	.351**	.329**
InsE1_M	.46	.15	.48	.16	.290**	.301**	.426**	.530**	1	.565**	.494**	.442**
InsE2_M	.50	.17	.62	.18	.181**	.321**	.359**	.487**	.606**	1	.536**	.476**
CoE1_M	.52	.15	.57	.16	.284**	.335**	.427**	.462**	.380**	.451**	1	.590**
CoE2_M	.61	.15	.54	.15	.277**	.306**	.394**	.501**	.367**	.429**	.539**	1

Note. American sample (US) n = 341; Chinese sample (CN) n = 345; correlations in the American sample were below the diagonal; correlations in the Chinese sample were above the diagonal; LME1_M, LME2_M, UseE1_M, UseE2_M, InsE1_M, InsE2_M, CoE1_M, CoE2_M refer to average scores obtained over the six ideas evaluated on each of the two DT subtests of Line Meanings, Uses, Instances, and Consequences respectively.

**$p < .01$.

Initial Reliability Analyses

The internal consistency coefficients were .709, .776, .752, .699 for Line Meanings, Uses, Instances, and Consequences respectively for the American group, and .740, .777, .719, .740 for the Chinese group, which were considered moderate reliability. In terms of the overall alpha coefficient over all eight subtests, the values were .854 for the American group, and .862 for the Chinese group, which indicated potential multidimensionality of the evaluative measures.

Confirmatory Factor Analysis

To answer the first question on the latent structure of the measures, CFA models with different number of factors were tested. Figure 5 displays the theoretical models to be tested for fluency scores. The one-factor model assumed the eight subtests measured a universal concept or construct of evaluative skill, whereas the four-factor model assumed distinctiveness of each kind of evaluative measures based on different types of DT tests (Runco et al., 2016). Table 19 shows the chi square statistic and fit indices for the two potential factor models when factor loadings were freely estimated in each cultural group as in the configural MI model. There was difference in chi square but not tested since they were not nested models. In terms of model fit, all indices suggested that the one-factor model showed a poor fit (e.g., CFI = .844), whereas the four-factor model fit the data well (e.g., CFI = .971). The results supported that the eight subtests provide information about four components of evaluative skill. Thus, the following steps of testing MI were based on the four-factor model.

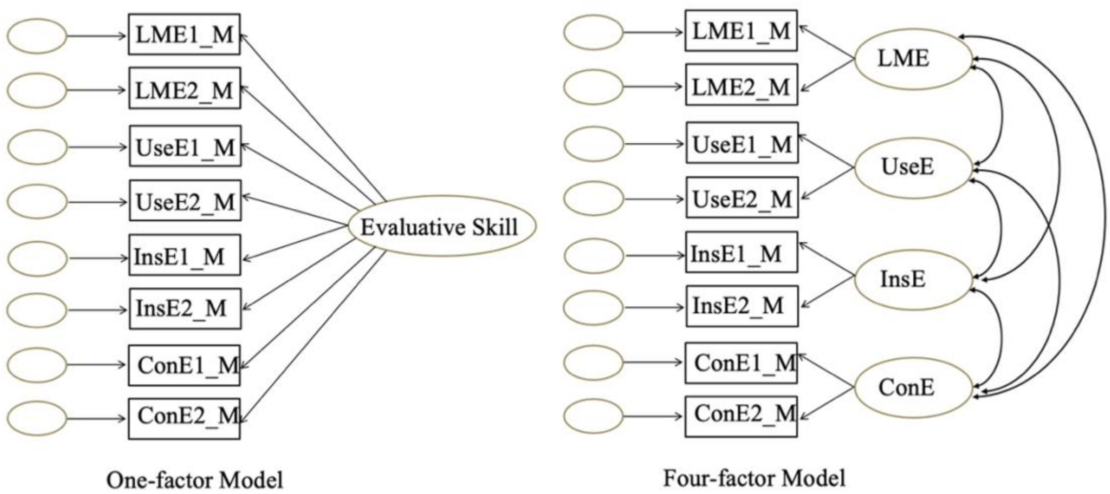


Figure 5. *Theoretical Models to be Tested: Evaluative Accuracy*

Note. LME1_M, LME2_M, UseE1_M, UseE2_M, InsE1_M, InsE2_M, CoE1_M, CoE2_M refer to average scores obtained over the six ideas evaluated on each of the two DT subtests of Line Meanings, Uses, Instances, and Consequences respectively. LME = Line Meanings Evaluative Accuracy; UseE = Uses Evaluative Accuracy; InsE = Instances Accuracy; ConE= Consequences Accuracy.

Table 19

Results of Model Comparison of One- and Four-Factor Models: Evaluative Accuracy

Model	No. of factors	χ^2	*df*	CFI	TLI	RMSEA	SRMR	$\Delta\chi^2$
Configural	One	378.053	40	.844	.782	.157	.070	
Configural	Four	90.606	28	.971	.942	.081	.032	287.447

Note. df = degree of freedom; CFI = Comparative fit index; TLI = Tucker Lewis index; RMSEA = Root mean square error of approximation; SRMR = Standardized root mean square residual.

Testing of Measurement Invariance

Parameter estimates based on the four-factor configural MI model are displayed in Table 20, and the model fit results for each step of MI tests are shown in Table 21. For the configural MI model, with the factor loading on the first item set at 1 while those on the second item freely estimated, results showed positive factor loadings estimated for each indicator-factor relationship in each cultural group. These estimates were significant; combined with the fit indices information indicating a good model fit (e.g., CFI = .971), these results together suggested the *configural invariance was supported* in the present analysis. In addition, the difference in CFI between the configural MI model and weak MI model was -.003, meeting the ΔCFI <= .01 criterion (Cheung & Rensvold, 2002) and supporting *the assumption of weak invariance*. However, the difference in CFI between the weak MI model and strong MI model was -.087, substantially exceeding the ΔCFI <= .01 criterion and suggesting *the violation of strong invariance*. Informed by modification indices based on the strong MI model, two subtests of Consequences were potential problem items, followed by two subtests of Instances, and two subtests of Uses. With the two intercepts of Consequences across the groups freely estimated (CFI = .909), the partial strong MI model was tested but not supported. Several other kinds of partial MI models were tested as well but they were not supported either.

Table 20

Configural MI Model Parameter Estimates on Evaluative Accuracy across Two Groups: Four Factors

	LME by		UseE by		InsE by		ConE by	
	LME1_M	LME2_M	UseE1_M	UseE2_M	InsE1_M	InsE2_M	ConE1_M	ConE2_M
American	@1	1.252	@1	1.075	@1	1.113	@1	1.040
Chinse	@1	0.977	@1	1.129	@1	0.923	@1	0.809

Note. LME = Line Meanings evaluative accuracy; UseE = Uses evaluative accuracy; InsE = Instances evaluative accuracy; ConsE = Consequences evaluative accuracy.

***$p < .001$.

Table 21

Fit Indices for Model Comparisons on Evaluative Accuracy across Two Groups: Four Factors

Model	χ^2	*df*	CFI	RMSEA	SRMR	Compare with	$\Delta\chi^2$	Δdf	ΔCFI
Configural	90.606	28	.971	.081	.032				
Weak	100.406	32	.968	.079	.044	configural	9.8*	4	-.003
Strong	294.060	36	.881	.145	.076	weak	193.654***	4	-.087

Note. *df* = degree of freedom; CFI = Comparative fit index; RMSEA = Root mean square error of approximation; SRMR = Standardized root mean square residual.

*$p < .05$; ***$p < .001$.

Next, item reduction was conducted to produce an invariant scale (e.g., Sturm et al., 2017; Williams & Gotham, 2021). Informed by the initial reliability analyses, the modification indices analyses above, and the nature of the eight subtests (i.e., every two subtests were based on a single type of DT test), the two items on both the Instances and Consequences test were removed and the group of nested models were fitted again. To confirm whether the two-factor model remained the better model, the one-factor and two-factor models (see Figure 6) were tested again, and results are shown in Table 22 when factor loadings were freely estimated in each cultural group as in the configural MI model. Again, results suggested a significantly better fit of the two-factor model (e.g., CFI = 1) than the one-factor model (e.g., CFI = .937). The results further supported that the two subtests selected provide information about two components of evaluative skill. Thus, the following steps of testing MI were based on the two-factor model.

Parameter estimates based on the two-factor configural MI model are displayed in Table 23, and the model fit results for each step of MI tests are shown in Table 24. For the configural MI model, with the factor loading on the first item set at 1 while those on the second item freely estimated, results showed positive factor loadings estimated for each indicator-factor relationship in each cultural group. These estimates were significant; combined with the fit indices information indicating a good model fit (e.g., CFI = 1), these results together suggested the *configural invariance was supported* in the present analysis. In addition, the difference in CFI between the configural MI model and weak MI model was .001, meeting the ΔCFI <= .01 criterion (Cheung & Rensvold, 2002) and supporting *the assumption of weak invariance*. Moreover, the difference in CFI between the weak MI model and strong MI model was -.011, marginally meeting the ΔCFI <= .01 criterion and suggesting *the satisfaction of strong invariance*.

Based on this strong MI model input, with the latent means in the American group set at zero while those in the Chinese group freely estimated, the latent mean comparison showed no significant difference between the two groups on the Line Meanings evaluation task; however, it showed higher evaluative scores of the Chinese sample on the Uses evaluation task (p <.001). Since this evaluative score refers to the discrepancy of respondents' ratings from the criterion values (i.e., lower values indicate high evaluative skill), this result indicated that the evaluative skill of the Chinese sample was lower than that of the American sample on evaluative tasks based on Uses tests.

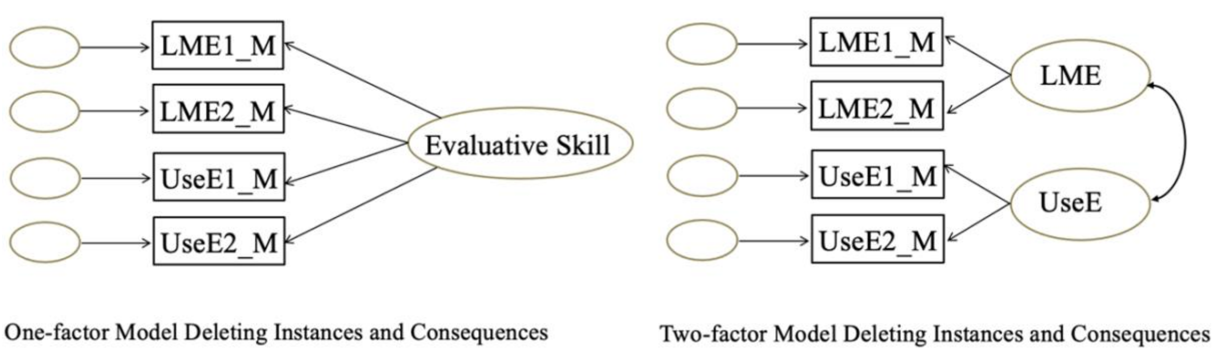


Figure 6. *Adjusted Theoretical Models to be Tested: Evaluative Accuracy*

Note. LME1_M, LME2_M, UseE1_M, UseE2_M refer to average scores obtained over the six ideas evaluated on each of the two DT subtests of Line Meanings, Uses, Instances, and Consequences respectively. LME = Line Meanings Evaluative Accuracy; UseE = Uses Evaluative Accuracy.

Table 22

Results of Model Comparison of One- and Two-Factor Models: Evaluative Accuracy

Model	No. of factors	χ^2	*df*	CFI	TLI	RMSEA	SRMR	$\Delta\chi^2$
Configural	One	62.799	4	.937	.812	.207	.043	
Configural	Two	0.324	2	1.000	1.000	.000	.003	62.475

Note. df = degree of freedom; CFI = Comparative fit index; TLI = Tucker Lewis index; RMSEA = Root mean square error of approximation; SRMR = Standardized root mean square residual.

Table 23

Configural MI Model Parameter Estimates on Evaluative Accuracy across Two Groups: Two Factors

	LME by		UseE by	
	LME1_M	LME2_M	UseE1_M	UseE2_M
American	@1	1.238***	@1	0.856***
Chinese	@1	1.049***	@1	1.053***

Note. LME = Line Meanings evaluative accuracy; UseE = Uses evaluative accuracy.

****p* < .001.

Table 24

Fit Indices for Model Comparisons on Evaluative Accuracy across Two Groups: Two Factors

Model	χ^2	*df*	CFI	RMSEA	SRMR	Compare with	$\Delta\chi^2$	Δdf	ΔCFI
Configural	.324	2	1.000	.000	.003				
Weak	4.656	4	.999	.022	.028	configural	4.332	2	.001
Strong	16.973	6	.988	.073	.039	weak	12.317**	2	-.011

Note. df = degree of freedom; CFI = Comparative fit index; RMSEA = Root mean square error of approximation; SRMR = Standardized root mean square residual.

***p* < .01.

The Individualism-collectivism Test

Initial Reliability Analyses

The internal consistency coefficients were .609, .696, .723, .654 for HI, VI, HC, and VC respectively for the American group, and .588, .671, .677, .682 for the Chinese group, which were considered moderate reliability. The Cronbach's alpha-if-item-deleted feature showed that removing Items 13 and 16 could result in significant increase in alpha value for the American group, and deleting Items 9 and 16 could result in significant increase in alpha value for the Chinese group. In terms of the alpha coefficient over all 16 items, the values were .682 for the American group and .766 for the Chinese group, which indicated potential multidimensionality of the individualism-collectivism test.

Confirmatory Factor Analysis

To answer the first question on the latent structure of the test, an initial CFA was conducted using all 16 items. Figure 7 (on the left) displays the theoretical model to be tested. This four-factor model assumed four dimensions of the individualism-collectivism test, which was suggested by the developers of this scale (Triandis & Gelfland, 1998). Factor loadings were freely estimated in each cultural group, and results showed that the model did not fit the data well (for the American group: $\chi^2 = 310.428$, $df = 98$, CFI = .821, RMSEA = .080, SRMR = .079; for the Chinese group: $\chi^2 = 426.088$, $df = 98$, CFI = .749, RMSEA = .099, SRMR = .080). Next, modification indices were examined. It suggested allowing items to load on other factors, for example allowing Item 10 to load on HI. Suggestions by modification indices are data driven, and when there is conflict between data and theories, the latter should prevail (Frazier & Youngstrom, 2007; Garcia-Barrera et al., 2011). In light of this, model modifications were not considered here, and instead, an Exploratory Factor Analyses (EFA) was conducted using Mplus based on four factors employing the Geomin rotation method.

Based on this EFA results, previous literature (Yao et al., 2012; Items 6, 9, 14 were dropped), and also the above mentioned initial reliability analyses showing the potential problem of Items 9, 13, and 16, the present study dropped seven items for both groups: Items 3, 6, 7, 9, 13, 14, and 16. The model on the right in Figure 7 displays the theoretical model to be tested using confirmatory factor analyses after removing the seven items. The revised model fit the data well (for the American group: $\chi^2 = 35.635$, $df = 21$, CFI = .972, RMSEA = .045, SRMR = .035; for the Chinese group: $\chi^2 = 26.147$, $df = 21$, CFI = .988, RMSEA = .027, SRMR = .035). Thus, the following steps of testing MI were based on the revised four-factor model.

Testing of Measurement Invariance

Parameter estimates based on the revised four-factor configural MI model are displayed in Table 25, and the model fit results for each step of MI tests are shown in Table 26. For the configural MI model, with the factor loading on the first item set at 1 while those on the second item freely estimated, results showed positive factor loadings estimated for each indicator-factor relationship in each cultural group. These estimates were significant; combined with the fit indices information indicating a good model fit (e.g., CFI = .980), these results together suggested *the configural invariance was supported* in the present analysis. In addition, the difference in CFI between the configural MI model and weak MI model was .002, meeting the ΔCFI <= .01 criterion (Cheung & Rensvold, 2002) and *supporting the assumption of weak invariance*. Moreover, the difference in CFI between the weak MI model and strong MI model was -.076, exceeding the ΔCFI <= .01 criterion and suggesting *the violation of strong invariance*. Because the aim of the present study was to examine the relationship between individualism-collectivism and creativity, the satisfaction of weak MI was satisfactory, and no further modifications to the scale were made.

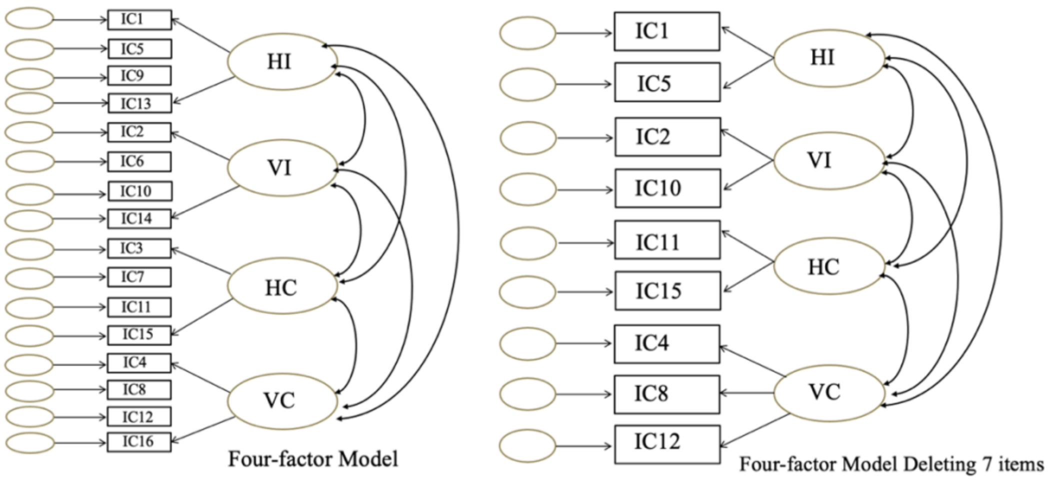


Figure 7. *Theoretical Models to be Tested: Individualism-collectivism*

Note. The model on the left refers to the four-factor model containing all 16 items, and the one on the right represents the four-factor model after deleting seven items. HI = Horizontal Individualism; VI = Vertical Individualism; HC = Horizontal Collectivism; VC = Vertical Collectivism.

Table 25

Configural MI Model Parameter Estimates on the Individualism-Collectivism Test across Two Groups

	HI by		VI by		HC by		VC by		
	IC1	IC5	IC2	IC10	IC11	IC15	IC4	IC8	IC12
American	@1	0.993***	@1	1.226***	@1	0.536***	@1	0.977***	1.629***
Chinese	@1	0.920***	@1	1.041***	@1	0.449***	@1	0.838***	1.336***

Note. HI = Horizontal Individualism; VI = Vertical Individualism; HC = Horizontal Collectivism; VC = Vertical Collectivism.

***p < .001.

Table 26

Fit Indices for Model Comparisons on the Individualism-Collectivism Test across Two Groups

Model	χ^2	*df*	CFI	RMSEA	SRMR	Compare with	$\Delta\chi^2$	Δdf	ΔCFI
Configural	61.782	42	.980	.037	.035				
Weak	63.974	47	.982	.032	.036	configural	2.192	5	.002
Strong	143.439	52	.906	.072	.051	weak	79.465***	5	-.076

Note. *df* = degree of freedom; CFI = Comparative fit index; RMSEA = Root mean square error of approximation; SRMR = Standardized root mean square residual.

***$p < .001$.

The Uncertainty Avoidance Test

Initial Reliability Analyses

The internal consistency coefficients over all six items were .904 for the American group, and .803 for the Chinese group, which were considered moderate to high reliability. The Cronbach's alpha-if-item-deleted feature showed no item could significantly increase the alpha value for either group.

Confirmatory Factor Analysis

To answer the first question on the latent structure of the test, an initial one-factor CFA was conducted using all six items. This model (see Figure 8 on the left) assumed the unitary dimension of the uncertainty avoidance test, which was suggested by previous works (Hofstede, 1980a). Factor loadings were freely estimated in each cultural group, and results showed that the model did not fit the data well (for the American group: $\chi^2 = 201.126$, $df = 9$, CFI = .865, RMSEA = .250, SRMR = .054; for the Chinese group: $\chi^2 = 135.462$, $df = 9$, CFI = .811, RMSEA = .202, SRMR = .069). Next, EFA was conducted with Mplus based on two factors employing the Geomin rotation method. Based on this EFA result, the present study dropped the cross-loaded Item 3 (i.e., I dislike unpredictable situations) for both groups. Based on the analysis on the content of each item, two factors can be extracted in terms of affective and cognitive aspects, referred to here as Emotional Response, and Dislike of Uncertainty. Figure 8 (on the right) displayed the theoretical models to be tested using confirmatory factor analyses. The revised model fit the data well (for the American group: χ^2 = 9.203, *df* = 4, CFI = .995, RMSEA = .062, SRMR = .019; for the Chinese group: χ^2 = 3.845, *df* = 4, CFI = 1, RMSEA = 0, SRMR = .013). Thus, the following steps of testing MI were based on the revised two-factor model.

Testing of Measurement Invariance

Parameter estimates based on the revised two-factor configural MI model are displayed in Table 27, and the model fit results for each step of MI tests are shown in Table 28. For the configural MI model, with the factor loading on the first item set at 1 while those on the second item freely estimated, results showed positive factor loadings estimated for each indicator-factor relationship in each cultural group. These estimates were significant; combined with the fit indices information indicating a good model fit (e.g., CFI = .997), these results together suggested *the configural invariance was supported* in the present analysis. In addition, the difference in CFI between the configural MI model and weak MI model was -.002, meeting the ΔCFI <= .01 criterion (Cheung & Rensvold, 2002) and *supporting the assumption of weak invariance*. Moreover, the difference in CFI between the weak MI model and strong MI model was -.005, additionally meeting the ΔCFI <= .01 criterion and suggesting *the assumption of strong invariance*.

Although not the focus of the present study, results of the latent mean comparison on uncertainty avoidance were obtained here to inform other conclusions of interest. Based on the strong MI model input, with the latent means in the American group set at zero while those in the Chinese group freely estimated, the latent mean comparison showed higher scores of the Chinese sample on both factors of Emotional Response and Dislike of Uncertainty ($p < .05$ and $p < .001$ respectively) compared to the American sample.

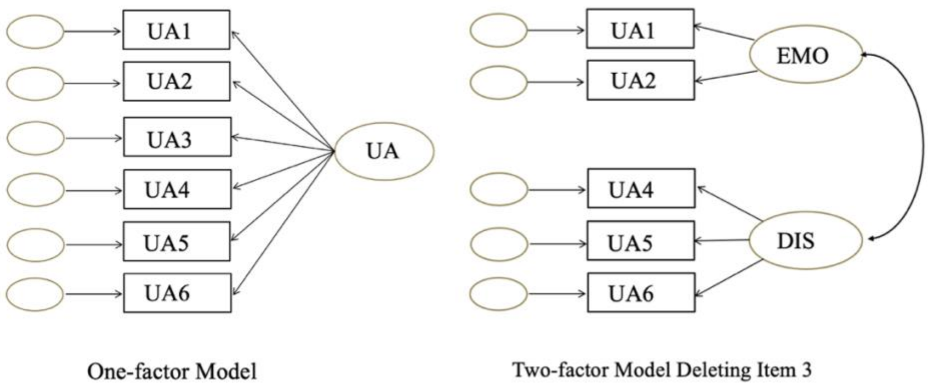


Figure 8. *Theoretical models to be tested: Uncertainty Avoidance*

Note. The model on the left refers to the one-factor model containing all 6 items, and the one on the right represents the two-factor model after deleting Item 3. UA1, UA2, UA3, UA4, UA5, UA6 represent the six items on the Uncertainty Avoidance Test.

Table 27

Configural MI Model Parameter Estimates on the Uncertainty Avoidance Test across Two Groups

	EMO by		DIS by		
	UA1	UA2	UA4	UA5	UA6
American	@1	0.995***	@1	1.608***	1.586***
Chinese	@1	1.331***	@1	1.470***	1.456***

Note. EMO = Emotional Response; DIS = Dislike of Uncertainty.

***$p < .001$.

Table 28

Fit Indices for Model Comparisons on the Uncertainty Avoidance Test across Two Groups

Model	χ^2	*df*	CFI	RMSEA	SRMR	Compare with	$\Delta\chi^2$	Δdf	ΔCFI
Configural	13.048	8	.997	.043	.016				
Weak	18.724	11	.995	.045	.037	configural	5.676	3	-.002
Strong	29.754	14	.990	.057	.045	weak	11.03*	3	-.005

Note. df = degree of freedom; CFI = Comparative fit index; RMSEA = Root mean square error of approximation; SRMR = Standardized root mean square residual.

$^*p < .05$.

Summary of MI Tests Results for Measures of all Constructs

In the above section, the measurement invariance for measures of all constructs were tested, and results (see Table 29 for summary) showed: (a) For DT fluency scores, configural and weak MI of the original four-factor model with all eight subtests were met, but strong MI was not; however, after deleting the two subtests of Instances, the strong MI of the revised three-factor model was satisfied. (b) For DT originality scores, configural MI of the original four-factor model with all eight subtests were met, but weak and strong MI were not; after deleting the two subtests of both Uses and Instances, the weak and strong MI of the revised two-factor model were satisfied. (c) For evaluative measures, configural and weak MI of the original four-factor model with all eight subtests were met, but strong MI was not; however, after deleting items based on the two subtests of both Instances and Consequences, strong MI of revised two-factor model was marginally satisfied. (d) For the individualism-collectivism test, the original four-factor model with 16 items did not fit the data well. After removing seven items, configural and weak MI of the revised 9-item four-factor model were met, but strong MI was not. (e) For the uncertainty avoidance test, the original one-factor model with six items did not fit the data well. After deleting one item, configural, weak and strong MI of the revised 5-item two-factor model were met.

Table 29

Summary of MI Tests Results for Measures of all Constructs

Construct	Original factor model	Revised factor model
Divergent thinking fluency	Original four-factor model: configural and weak MI met	Revised three-factor model (after removing the two subtests of Instances): configural, weak, and strong MI met
Divergent thinking originality	Original four-factor model: configural MI met	Revised three-factor model (after removing the two subtests of Instances): configural and weak MI met; Revised two-factor model (after removing the two subtests of Uses and Instances): configural, weak and strong MI met
Evaluative skill	Original four-factor model: configural and weak MI met	Revised two-factor model (after removing items based on the two subtests of Instances and Consequences): configural and weak MI met, strong MI marginally satisfied
Individualism-collectivism	Original four-factor model did not fit the data well	Revised four-factor model (after deleting seven items): configural and weak MI met
Uncertainty Avoidance	Original one-factor model did not fit the data well	Revised two-factor model (after deleting one item): configural, weak and strong MI met

SEM Results for Relationships between all Variables

SEM Results with Fluency as DT Output

The present study aimed to investigate the relationships between individualism-collectivism, uncertainty avoidance, divergent thinking, and evaluative skill. The satisfaction of weak invariance is required to explore relationships between constructs across groups (Chen & West, 2008; Rusticus & Hubley, 2006). The MI tests results above showed that the measures of all constructs of interest met weak invariance based on either the original or revised model. The revised model (after removing items) needed to reach strong MI would not be considered here if weak MI of the original model was already supported. Specifically, the original four-factor model (Figure 1-left) for DT fluency, the revised three-factor model (Figure 4-left) for DT originality, the original four-factor model (Figure 5-left) for evaluative skill, the revised four-factor model (Figure 7-right) for the individualism-collectivism test and the revised two-factor model (Figure 8-right) for the uncertainty avoidance test were considered legitimate measurement models to enter the structural models.

The theoretical relationships in the structural models to be tested were as follows: individualism-collectivism (four latent factors) and uncertainty avoidance (two latent factors) served as the predictor variables; divergent thinking (four latent factors for fluency or three latent factors for originality) and evaluative skill (four latent factors) served as the outcome variables. Table 30 shows the model fit results when fluency was used as the DT output, which indicated that the model fit was good for both the American and Chinese groups. Figure 9 and Figure 10 display the model results with significant coefficients for the American and Chinese group respectively (to facilitate perusal of the figures, solid lines indicate significant effects, whereas insignificant effects are not displayed). Specifically, for the American group, path analysis showed that HC positively predicted DT fluency based on Line Meanings, and VC negatively predicted DT fluency based on Instances. For the Chinese group, it showed that VC negatively predicted DT fluency based on Lines Meanings and Uses, and positively predicted evaluative scores (i.e., negatively predicted evaluative accuracy since the score refers to discrepancy from the criterion) based on Line Meanings tasks.

For the association between DT fluency and evaluative scores, for the American group, results showed a positive relation (i.e., a negative relation between DT fluency and evaluative accuracy) in general when the evaluation task was based on Line Meanings and Uses; however, there was a negative link between the two (i.e., a positive relation between DT fluency and evaluative accuracy) when the evaluation task was based on Instances. For the Chinese group, similar to the American group, there was a positive association between DT fluency and evaluative scores (i.e., a negative association between DT fluency and evaluative accuracy) in general, but without any negative relation.

Table 30

Fit Indices for SEM models for all Variables: Fluency as Divergent Thinking Output

Model	χ^2	*df*	CFI	TFI	RMSEA	SRMR	AIC	BIC
American	462.395***	314	.964	.951	.037	.038	25313.696	26007.267
Chinese	464.092***	314	.958	.942	.037	.041	23439.466	24135.148

Note. *df* = degree of freedom; CFI = Comparative fit index; TFI = Tucker Lewis index; RMSEA = Root mean square error of approximation; SRMR = Standardized root mean square residual; AIC = Akaike's information criteria; BIC = Bayesian information criteria.

***$p < .001$.

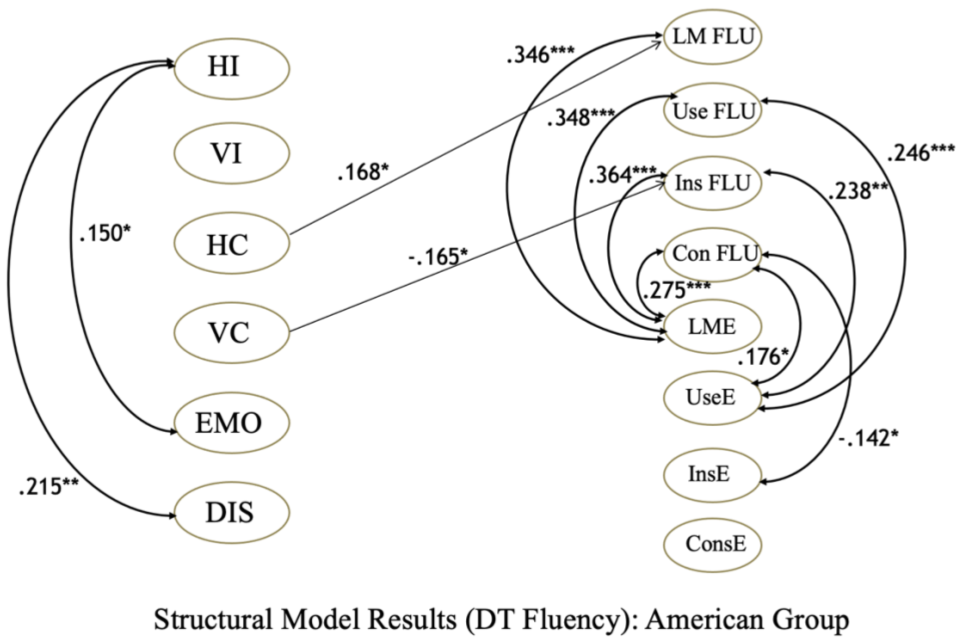


Figure 9. *SEM Model Results for the American Group: Fluency as Divergent Thinking Output*

Note. All coefficient estimates are standardized. Solid lines indicate significant effects, whereas insignificant effects are not displayed. In addition, relations between latent variables of the same construct are not displayed, specifically: Those between HI, VI, HC, and VC; those between EMO and DIS; those between LM FLU, Use FLU, Ins FLU, and Con FLU; those between LME, UseE, InsE, and ConsE. Special note: The path from VC to Instances fluency was marginally significant, with $p = .054$.

HI = horizontal individualism; VI = vertical individualism; HC = horizontal collectivism; VC = vertical collectivism; EMO = emotional response; DIS = dislike of uncertainty; LM FLU = Line Meanings fluency; Use FLU = Uses fluency; Ins FLU = Instances fluency; Con FLU = Consequences fluency; LME = Line Meanings evaluative accuracy; UseE = Uses evaluative accuracy; InsE = Instances evaluative accuracy; ConsE = Consequences evaluative accuracy.

*p < .05; **p < .01; ***p < .001.

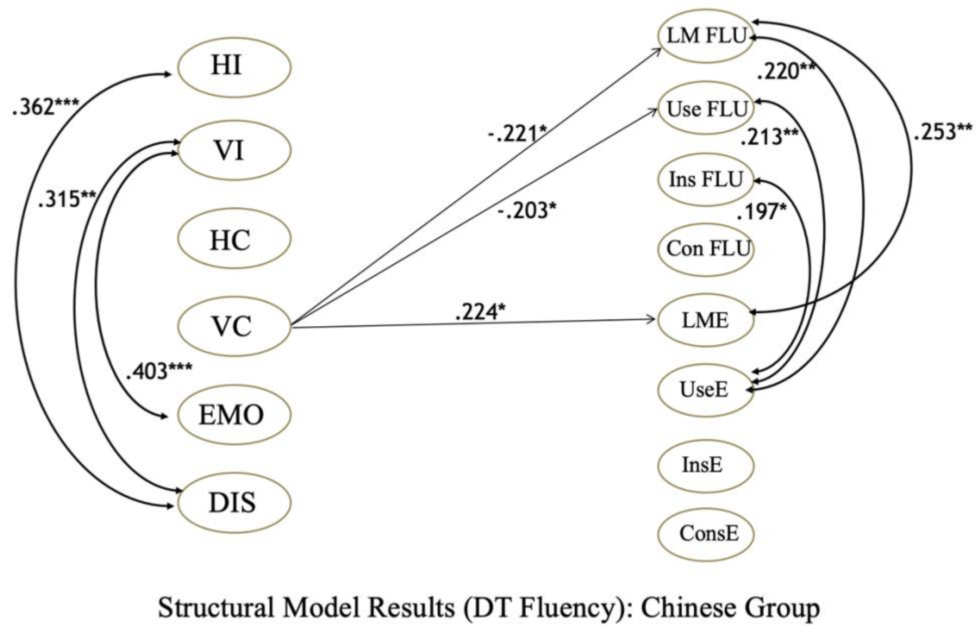


Figure 10. *SEM Model Results for the Chinese Group: Fluency as Divergent Thinking Output*

Note. All coefficient estimates are standardized. Solid lines indicate significant effects, whereas insignificant effects are not displayed. In addition, relations between latent variables of the same construct are not displayed, specifically: Those between HI, VI, HC, and VC; those between EMO and DIS; those between LM FLU, Use FLU, Ins FLU, and Con FLU; those between LME, UseE, InsE, and ConsE. Special note: The path from VC to LME was marginally significant, with p = .055.

HI = horizontal individualism; VI = vertical individualism; HC = horizontal collectivism; VC = vertical collectivism; EMO = emotional response; DIS = dislike of uncertainty; LM FLU = Line Meanings fluency; Use FLU = Uses fluency; Ins FLU = Instances fluency; Con FLU = Consequences fluency; LME = Line Meanings evaluative accuracy; UseE = Uses evaluative accuracy; InsE = Instances evaluative accuracy; ConsE = Consequences evaluative accuracy.

*p < .05; **p < .01; ***p < .001.

SEM Results with Originality as DT Output

Accordingly, Table 31 shows the model fit results when originality was used as the divergent thinking output, which indicated that the model fit was good for both the American and Chinese groups. Figure 11 and Figure 12 display the model results with significant coefficient estimates for American and Chinese group respectively (to facilitate perusal of the figures, solid lines indicate significant effects, whereas insignificant effects are not displayed). Specifically, for the American group, path analysis showed that VC negatively predicted DT originality based on Consequences. For the Chinese group, it showed that VC negatively predicted DT originality based on Lines Meanings and Uses, and positively predicted evaluative scores (i.e., negatively predicted evaluative accuracy because the score refers to discrepancy from the criterion) when the evaluation task was based on Line Meanings. For the association between DT originality and evaluative scores, for both the American and Chinese groups, results showed a positive relation (i.e., a negative relation between DT originality and evaluative accuracy) in general when the evaluation task was based on Line Meanings and Uses.

Table 31

Fit Indices for SEM models for all Variables: Originality as Divergent Thinking Output

Model	χ^2	*df*	CFI	TFI	RMSEA	SRMR	AIC	BIC
American	391.157***	272	.963	.949	.036	.037	21319.579	21940.344
Chinese	376.838***	272	.959	.942	.033	.041	19918.318	20540.973

Note. df = degree of freedom; CFI = Comparative fit index; TFI = Tucker Lewis index; RMSEA = Root mean square error of approximation; SRMR = Standardized root mean square residual; AIC = Akaike's information criteria; BIC = Bayesian information criteria.

***$p < .001$.

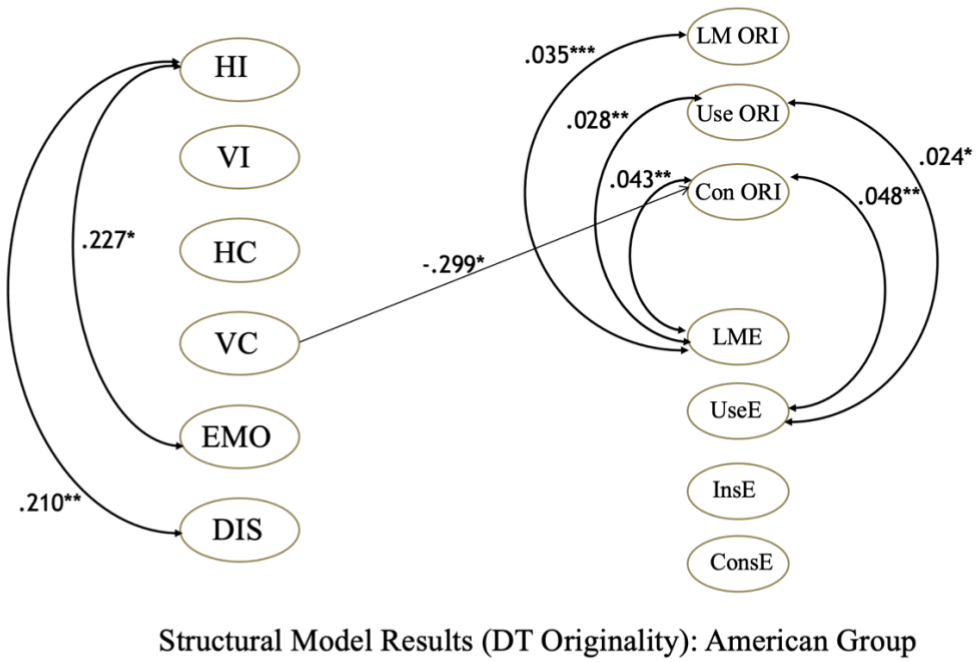


Figure 11. *SEM Model Results for the American Group: Originality as Divergent Thinking Output*

Note. All coefficient estimates are standardized. Solid lines indicate significant effects, whereas insignificant effects are not displayed. In addition, relations between latent variables of the same construct are not displayed, specifically: Those between HI, VI, HC, and VC; those between EMO and DIS; those between LM ORI, Use ORI, and Con ORI; those between LME, UseE, InsE, and ConsE.

HI = horizontal individualism; VI = vertical individualism; HC = horizontal collectivism; VC = vertical collectivism; EMO = emotional response; DIS = dislike of uncertainty; LM FLU = Line Meanings originality; Use FLU = Uses originality; Con FLU = Consequences originality; LME = Line Meanings evaluative accuracy; UseE = Uses evaluative accuracy; InsE = Instances evaluative accuracy; ConsE = Consequences evaluative accuracy.

*p < .05; **p < .01; ***p < .001.

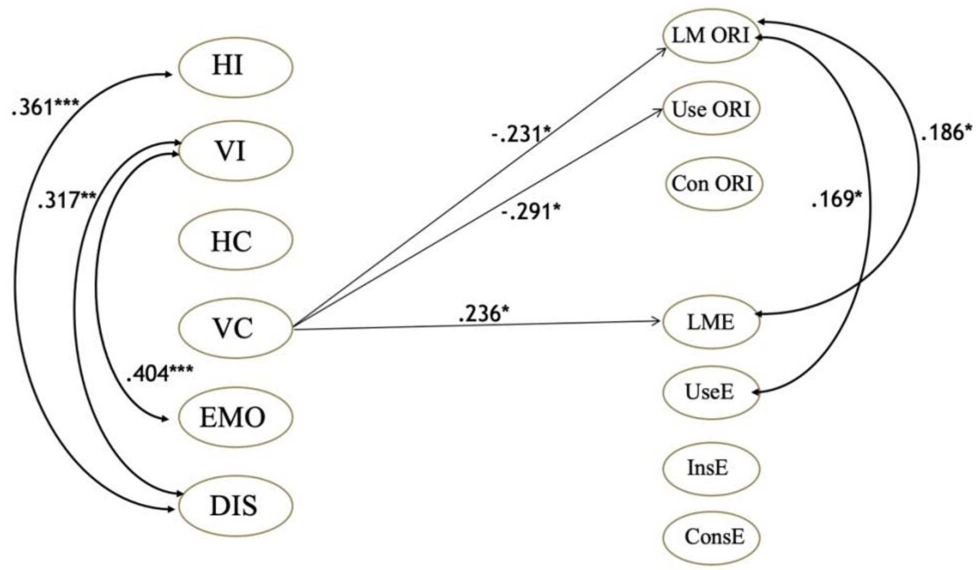


Figure 12. *SEM Model Results for the Chinese Group: Originality as Divergent Thinking Output*

Note. All coefficient estimates are standardized. Solid lines indicate significant effects, whereas insignificant effects are not displayed. In addition, relations between latent variables of the same construct are not displayed, specifically: Those between HI, VI, HC, and VC; those between EMO and DIS; those between LM ORI, Use ORI, and Con ORI; those between LME, UseE, InsE, and ConsE.

HI = horizontal individualism; VI = vertical individualism; HC = horizontal collectivism; VC = vertical collectivism; EMO = emotional response; DIS = dislike of uncertainty; LM FLU = Line Meanings originality; Use FLU = Uses originality; Con FLU = Consequences originality; LME = Line Meanings evaluative accuracy; UseE = Uses evaluative accuracy; InsE = Instances evaluative accuracy; ConsE = Consequences evaluative accuracy.

*p < .05; **p < .01; ***p < .001.

CHAPTER 5

DISCUSSION

This chapter discusses the results in the following order: divergent thinking tests, evaluative measures, the individualism-collectivism test, and the uncertainty avoidance test. For measures of each construct, it discusses the latent structure first, then the measurement invariance tests results, followed by latent mean comparisons results. Next, it discusses the results for the relationships between all variables: the relationship between cultural orientation and creativity, followed by that between DT and evaluative skill. Last, it draws a conclusion and discusses limitations of the present study.

Divergent Thinking Tests

Latent Structure

The present study aimed to investigate the latent factor structure and MI of four DT tests (Line Meanings, Uses, Instances, and Consequences) across American and Chinese cultural groups. First, the four-factor model showed a good fit for both groups on the fluency as well as originality scores, which is consistent with previous views that different types of DT measures are not equivalent, and there is no universal concept (or factor) of DT (Reiter-Palmon et al., 2019; Runco et al., 2016). The present study provided psychometric evidence and confirmed the argument by documenting four different latent factors based on different types of DT tasks, suggesting that scores from different DT measures may not be redundant and can provide unique information. This finding is also consistent with the recent MI study based on two DT tests (Guo et al., 2021). Therefore, as suggested by previous researchers (Long & Plucker, 2015; Reiter-

Palmon et al., 2019; Runco et al., 2016), various types of DT tests should be used in creativity research to generate a broader and more generalizable conclusion in terms of DT ability.

Measurement Invariance

In terms of measurement invariance, for DT fluency, based on the four-factor model with all eight subtests, different levels of MI tests using multi-group CFA analyses showed that configural invariance and weak invariance were satisfied, whereas strong invariance was violated. First, the assumption of configural MI indicated that the same four-factor latent structure for DT was assessed across American and Chinese groups, that is, “the same indicators are associated with the same factors for both groups” (Rensvold & Cheung, 2001, p. 29). Second, the satisfaction of weak MI indicated that the relationships between factors underlying DT or between DT and other external variables of interest were comparable, that is, “the measure may be used to examine structural relationships or correlations between the construct of interest and other constructs across groups” (Rusticus & Hubley, 2006, p. 828). Third, the violation of strong MI indicated a problem in comparing observed mean differences across groups. In this case, researchers should be cautious because it would be misleading to attribute observed mean differences to latent construct differences. As suggested by Thompson and Green (2006), at least partial strong MI should be satisfied for meaningful latent mean comparisons. Further, with the removal of two subtests of Instances, the revised three-factor model met strong MI. It indicated that latent mean comparison can be conducted across American and Chinese individuals with this three-factor model, and the Instances tests may be more culturally dependent in terms of DT fluency. It is possible that people living in different locations have different historical and social heritages (Cheung et al., 2016; Hartley et al., 2016) and thus may be substantively different in finding different examples (e.g., for things that are square or can move one wheels). Last, the recent MI study (Guo et al., 2021) across American and Chinese individuals based on DT tasks of Line Meanings and Real-world Problems supported weak MI of the two tests cross-culturally, but only partial strong MI of Real-real Problems fluency was satisfied, whereas that of Line Meanings still failed to be established. The present study also employed Line Meanings tasks with the result supporting the assumption of strong MI. Possible reasons include that the tasks in the present study used different items (i.e., figures) from the previous study; the respondents of this study and the previous one are from different locations (e.g., for Chinese participants, central China in the present study vs. eastern China in the previous one), and thus, factors such as SES may play a role.

In the case of originality, based on the four-factor model with all eight subtests, different levels of MI tests using multi-group CFA analyses showed satisfied configural invariance, but violated weak invariance. Similar to the case of fluency, the assumption of configural MI indicated that the same four-factor latent structure for DT was assessed across American and Chinese groups. Different from that of fluency, the violated weak invariance indicated that the relationships between factors underlying DT or between DT and other external variables of interest were not comparable (Rusticus & Hubley, 2006). Further, with the removal of two subtests of Instances, the revised three-factor model met weak MI, but strong MI was still not satisfied. This finding indicated that the relationships between factors underlying DT or between DT and other external variables of interest (e.g., individualism-collectivism) were comparable across the two groups based on this three-factor model. Next, with the removal of the two subtests of both Uses and Instances, the revised two-factor model finally met strong MI, legitimately permitting latent mean comparisons across the two cultural groups (Thompson & Green, 2006). In terms of the specific DT tests removed, it indicated that Uses and Instances are more culturally dependent. When compared to the MI tests results on fluency, it suggested that different DT indices (e.g., fluency, originality) may show different psychometric properties. The reason may be that originality depends more on specific cultures and different cultures may perceive originality in a different manner (Karwowski, 2016). For example, in term of the concept of creativity, individuals from the West tend to emphasize originality more while those from the East pay more attention to appropriateness (Adair & Xiong, 2018; Morris & Leung, 2010; Xie & Paik, 2019).

Latent Mean Comparisons

Results of latent mean comparisons on fluency and originality scores were similar. Specifically, based on the three-factor strong MI model (with Line Meanings, Uses, and Consequences) input, it showed higher DT fluency of American individuals compared to their Chinese counterparts; based on the two-factor strong MI model (with Line Meanings and Consequences) input, it showed higher DT originality of American individuals compared to Chinese peers. This is consistent with previous studies supporting the higher level of DT performance of American individuals compared to Chinese peers based on observed mean comparisons (e.g., Deng et al., 2016; Jellen & Urban, 1989). The present study provided more statistically robust evidence via latent mean comparisons (Barbot, 2019; Yuan & Bentler, 2006), and further clarified that not all DT tests (e.g., Instances) are comparable among those two cultural groups, and even the same DT test (e.g., Uses) may display different psychometric properties in terms of different DT indices (fluency, originality) among different cultural groups.

In sum, the results indicated that it is psychometrically valid to conduct latent mean comparisons based on the three-factor model (with Line Meanings, Uses, and Consequences) on fluency, and the two-factor model (with Line Meanings and Consequences) on originality. It is also valid to examine the relationship between DT fluency and other variables (e.g., individualism vs. collectivism) cross-culturally based on the four-factor model (with Line Meanings, Uses, Instances, and Consequences), and examine the relationship between DT originality and other variables (e.g., uncertainty avoidance) cross-culturally based on the three-factor model (with Line Meanings, Uses, and Consequences).

Evaluative Measures

Latent Structure

The present study aimed to investigate the latent factor structure and MI of four evaluative measures based on four different DT tests (Line Meanings, Uses, Instances, and Consequences) across American and Chinese cultural groups. First, the four-factor model showed a good fit for both groups on evaluative scores, which is consistent with previous views that evaluation tasks based on different types of DT tests may not equivalent. Specifically, Runco and Dow (2004) showed that participants had worse evaluative skill on the evaluation task based on Consequences compared to that based on Uses and Pattern Meanings. The present study confirmed the argument by documenting four different latent factors based on different types of evaluation tasks, suggesting that scores from different evaluative measures may not be redundant and can provide unique information. Therefore, various evaluative measures (i.e., ideas based on different DT tests) should be used to generate a more generalizable conclusion in terms of evaluative skill.

Measurement Invariance

In terms of measurement invariance, based on the four-factor model with all eight subtests, different levels of MI tests using multi-group CFA analyses showed satisfied configural invariance and weak invariance, and violated strong invariance. First, the assumption of configural MI indicated that the same four-factor latent structure for evaluative skill was assessed across American and Chinese groups. Second, the satisfaction of weak MI indicated that the relationships between factors underlying evaluative skill or between evaluative skill and other external variables of interest (e.g., individualism-collectivism) were comparable (Rusticus & Hubley, 2006). Third, the violation of strong MI indicated a problem when comparing observed mean differences across groups. In this case, researchers should be cautious because it would be misleading to attribute observed mean differences to latent construct differences. As suggested by Thompson and Green (2006), strong MI, or at least partial strong MI should be satisfied for meaningful latent mean comparisons. Further, with the removal of two subtests of both evaluation tasks based on Instances and Consequences, the revised two-factor model marginally met strong MI. It indicated that latent mean comparison can be conducted across American and Chinese individuals with this two-factor model, and evaluation tasks based on the Instances and Consequences tests may be more culturally dependent in terms of evaluative skill. It is possible that people living in different locations have different historical and social heritages (Cheung et al., 2016; Hartley et al., 2016) and thus may be substantively different in evaluating ideas based on the Instances and Consequences tasks. In contrast, the evaluation tasks based on the other two DT tests (Line Meanings, Uses) may be more culturally independent, indicating that they may be more applicable and valid tools in cross-cultural contexts.

Latent Mean Comparisons

The finding from latent mean comparisons based on the two-factor model (with evaluation tasks based on Line Meanings and Uses) supported higher evaluative skill of American respondents compared to their Chinese peers on the Uses evaluation task, although no difference emerged in that on the Line Meanings evaluation task. To my knowledge, this is the first study investigating evaluative skill cross-culturally, and the result supported that Americans may be better in judging creativity of ideas, especially on ideas based on the Uses test. Keeping in mind that this finding should be approached with caution because strong MI was only marginally satisfied, it suggested that compared to the results on DT performance (i.e., higher levels of American individuals on three DT tests in terms of fluency, and on two DT tests in terms of originality), people from the two cultural groups are more similar in evaluative skill. It is possible that Chinese individuals put more emphasis on appropriateness compared to American peers (Karwowski, 2016; Nijstad et al., 2010; Xie & Paik, 2019), which leads to their ignorance of the original aspect of ideas, thus they have worse evaluative skill.

In sum, the results indicated that it is psychometrically valid to conduct latent mean comparisons based on the two-factor model (with evaluation tasks based on Line Meanings and Uses) in terms of evaluative skill. It is also valid to examine the relationship between evaluative skill and other variables (e.g., individualism vs. collectivism) cross-culturally based on the four-factor model (with evaluation tasks based on Line Meanings, Uses, Instances, and Consequences).

The Individualism-collectivism Test

Latent Structure

The present study aimed to investigate the latent factor structure and MI of the individualism-collectivism test across American and Chinese cultural groups. First, the original four-factor model with 16 items did not show a good fit for both groups. This indicated that the original four-factor framework proposed by the developers (Triandis & Gelfland, 1998) may not be valid when assessing individualism-collectivism among American and Chinese individuals, with some items cross-loading on other factors, or not applicable to both groups. For example, for Item 6 (i.e., Winning is everything), results showed that besides loading on VI, it loads on VC as well, that is, this item is not a good indicator for VI. With the removal of seven items, the revised four-factor model showed a good fit. Therefore, partly consistent with previous research (Yao et al., 2012), item reduction needs to be done before the measure can validly assess the construct of interest.

Measurement Invariance

In terms of measurement invariance, based on the revised four-factor model with nine items, different levels of MI tests using multi-group CFA analyses showed satisfied configural invariance and weak invariance, and violated strong invariance. First, the assumption of configural MI indicated that the same four-factor latent structure for individualism-collectivism was assessed across American and Chinese groups. Second, the satisfaction of weak MI indicated that the relationships between factors underlying individualism-collectivism or between individualism-collectivism and other external variables of interest (e.g., evaluative skill) were comparable (Rusticus & Hubley, 2006). Third, the violation of strong MI indicated the problem in comparing observed mean differences across groups. In this case, researchers should be cautious because it would be misleading to attribute observed mean differences to latent construct differences. As suggested by Thompson and Green (2006), strong MI, or at least partial strong MI should be satisfied for meaningful latent mean comparisons.

In sum, the results indicated that it is psychometrically valid to examine the relationship between individualism-collectivism and other variables (e.g., divergent thinking, evaluative skill) within each cultural group, as well as compare the direction and strength of this relationship cross-culturally (Chen & West, 2008).

The Uncertainty Avoidance Test

Latent Structure

The present study aimed to investigate the latent factor structure and MI of the uncertainty avoidance test across American and Chinese cultural groups. First, the original one-factor model with six items did not show a good fit for both groups. This indicated that the original one-factor framework proposed by the developer (Hofstede, 1980a) may not be valid when assessing uncertainty avoidance among American and Chinese individuals. EFA analyses results showed one item (Item 3: I dislike unpredictable situations) cross-loaded on two factors. With the removal of this item, the revised two-factor model showed a good fit. Therefore, partly consistent with previous research (Jung, 2002), item reduction needs to be done before the measure can validly assess the construct.

Measurement Invariance

In terms of measurement invariance, based on the revised two-factor model with five items, different levels of MI tests using multi-group CFA analyses showed satisfied configural, weak and strong invariance. The assumption of configural and weak MI indicated that not only the latent structure of the construct (i.e., uncertainty avoidance) being examined was similar, but also the strength of the corresponding relationship between latent factors and individual indicators was the same across the two cultures. In addition, the assumption of strong MI indicated that the intercepts of indicator items in different groups were similar, suggesting the validity in comparing means in the two cultural groups (Thompson & Green, 2006).

In sum, the results indicated that based on the two-factor model it is psychometrically valid to examine the relationship between uncertainty avoidance and other variables (e.g., divergent thinking, evaluative skill) within each cultural group, as well as compare the direction and strength of this relationship cross-culturally (Chen & West, 2008). Directly comparing the means of uncertainty avoidance in two groups can also be conducted, although it was not the focus of the present study.

Relationships Between all Variables

The present study aimed to investigate the relationships between individualism-collectivism, uncertainty avoidance, divergent thinking, and evaluative skill. The satisfaction of weak invariance is required to explore relationships between constructs across groups (Chen & West, 2008; Rusticus & Hubley, 2006). Based on interpretations above, the original four-factor model (with Line Meanings, Uses, Instances, and Consequences) for DT fluency, the revised three-factor model (with Line Meanings, Uses, and Consequences) for DT originality, the original four-factor model for evaluative skill (with evaluation tasks based on Line Meanings, Uses, Instances, and Consequences), the revised four-factor model for the individualism-collectivism test (with nine items) and the revised two-factor model (with five items) for the uncertainty avoidance test were considered valid measurement models to enter the structural models. The theoretical relationships to be tested were as follows: individualism-collectivism (four latent factors) and uncertainty avoidance (two latent factors) served as the predictor variables; divergent thinking (four latent factors for fluency or three latent factors for originality) and evaluative skill (four latent factors) served as the outcome variables.

Cultural Orientation and Creativity

The present study investigated two conceptualizations of cultural orientation (individualism-collectivism, uncertainty avoidance), and two components of creativity (DT, evaluative skill). With fluency as the DT output, SEM analyses showed that for the American group, horizontal collectivism (HC) positively predicted DT fluency based on Line Meanings, and vertical collectivism (VC) negatively predicted DT fluency based on Instances. For the Chinese group, it showed that VC negatively predicted DT fluency based on Lines Meanings and Uses, and negatively predicted evaluative accuracy based on Line Meanings tasks. Taken together, the results provided evidence of (a) the negative effect of VC on DT fluency across different DT tasks (although not all tasks) and across cultures, (b) the positive effect of HC on DT fluency in the American group only, and (c) the negative effect of VC on evaluative skill in the Chinese group only. First, the consistent negative effect of VC on DT fluency was consistent with previous views that collectivism focuses on interdependence and conformity, which may hinder unique idea generation (Brew & Chen, 2007; Ng, 2001). Recall here that VC refers to the belief that people not only have common goals and are dependent, but also are distinct from each other in status, which is different from HC, which holds that people are similar to each other. As noted by Triandis and Gelfand (1998, p. 119), "The most important attributes that distinguish among different kinds of individualism and collectivism are the relative emphases on horizontal and vertical social relationships." The present finding further clarified that not all kinds of collectivism, but only vertical collectivism (i.e., people are different in status) negatively predicts DT fluency. Further, consistent with the finding of a previous study (Yao et al., 2012) investigating Chinese employees, the present study showed that HC positively predicts DT in the American group. It is possible that not collectivism itself, but the inequality of people's status makes them unable to freely express their opinions and thus inhibits idea generation, whereas equality can promote the ideational process. To some extent, this is consistent with previous researchers' (e.g., Oyserman, 2006; Shavitt et al., 2006) views that the vertical vs. horizontal dimension is similar to the concept of power distance proposed by Hofstede (1980a), and larger power distance may lead to inhibition on proposing novel ideas (Erez & Nouri, 2010; Morris & Leung, 2010).

With originality as the DT output, SEM analyses showed that for the American group, VC negatively predicted DT originality based on Consequences. For the Chinese group, it showed that VC negatively predicted DT originality based on Lines Meanings and Uses, and negatively predicted evaluative accuracy when the evaluation task was based on Line Meanings. Similar to the results with fluency as the DT output, those results also showed the consistent negative effect of VC on DT, suggesting the consistent impact of VC on different indices of divergent thinking. It is possible the inequality of people's status not only makes them unable to freely express their opinions and thus inhibits the generation of ideas in number, but also the production of original or unique ideas in quality (Erez & Nouri, 2010; Morris & Leung, 2010). Last, based on the model with either fluency or originality as the DT output, the negative effect of VC on evaluative skill in the Chinese group suggested that this vertical dimension not only inhibits the ideational process but also the evaluative process. It is possible that in collectivistic cultures people tend to conform (Brew & Chen, 2007; Kim, 2009; Ng, 2001), and the perceived different status of people increases the tendency, thus they do not appreciate the value of creative ideas to avoid potential confrontations.

Divergent Thinking (DT) and Evaluative Skill

With fluency as the DT output, SEM analyses showed that for the American group, there was a negative link between DT fluency and evaluative accuracy in general when the evaluation task was based on Line Meanings and Uses; however, there was a positive relation between DT fluency and evaluative accuracy when the evaluation task was based on Instances. For the Chinese group, similar to the American group, there was a negative association between DT fluency and evaluative accuracy in general. It is possible that DT belongs to ideational processes whereas evaluative skill is analytic, thus they are independent of each other (Grohman et al., 2006; Sternberg & Lubart, 1995). The consistent evidence of a negative link across cultures suggested that divergent thinking can be negatively affected by personal evaluations (Runco & Dow, 2004). In contrast, the evidence of a positive link indicated that the relationship between DT and evaluative skill may be moderated by the type of evaluation tasks used. People have different evaluative accuracy on different evaluation tasks (Runco & Dow, 2004), and this may affect the link between DT and evaluative skill. Previous studies on DT and evaluative skill also showed inconsistent results, with some supporting a positive link (Benedek et al., 2016; Runco, 1991) whereas others suggesting a negative one (Grohman et al., 2006; Guo et al., 2019). In this case, it is possible that some moderators (e.g., the type of DT tests, the type of evaluation tasks, the method of measuring evaluative accuracy, evaluation instructions, interpersonal vs. intrapersonal evaluation) may have played a role in affecting the relationship. For example, although the relationship was inconsistent in interpersonal evaluation, there appears a tendency that DT positively correlated with intrapersonal evaluation skill (e.g., Grohman et al., 2006; Runco & Smith, 1992). It is possible that in intrapersonal evaluation (i.e., evaluating one's own ideas), people are familiar with their own ideational process and understand the pathway of ideas better, whereas in interpersonal evaluation (i.e., evaluating others' ideas), people may need extra information for accurate idea judgement (Charles & Runco, 2001). Thus, future studies can further investigate the issue when people are evaluating their own ideas, and by more systematically varying the types of tasks, measuring methods, and instructions used to assess these constructs.

With originality as the DT output, SEM analyses showed that for both the American and Chinese groups, there was a negative relation between DT originality and evaluative accuracy in general when the evaluation task was based on Line Meanings and Uses. Those results further confirmed the negative link between DT and evaluative skill regardless of DT indices or outputs.

Conclusion

The present study is one of the first to investigate the differences in evaluative skill across the Eastern and Western cultures via latent mean comparisons, which can draw more statistically robust conclusions. The findings contribute to the understanding that there are cross-cultural differences in people's evaluative skill, and the pattern is different from that on divergent thinking. Specifically, it provides evidence that Westerners have higher performance on evaluative skill based on some evaluation tasks in addition to popularly researched DT. It also suggests that DT and evaluative skill (when people are evaluating others' ideas) are independent of each other or even negatively correlated. In this case, it implies that educators should raise the concern of promoting students' evaluative skill in addition to DT, particularly for those from collectivistic cultures. For example, students may need to be provided more opportunities to practice evaluating others' ideas.

Further, although previous views suggested that collectivism inhibited idea generation (Brew & Chen, 2007; Ng, 2001), the present study further clarified that not all kinds of collectivism, but only vertical collectivism (i.e., people are different in status) negatively predicts both DT and evaluative skill. This suggests that sharing common goals and being dependent on each other in collectivistic cultures not necessarily hinders creativity, but the unequal status between people plays an important role. The implication in educational settings is that a more equal teacher-student relationship might be helpful for the development of individuals' creativity in the classroom.

In addition, it is also one of the few studies that further investigated cross-cultural differences on DT via latent mean comparisons. First, the examination of the latent structure for both DT and evaluative skill suggested the multidimensional nature of both constructs. It implies that future researchers should employ more than one creativity tests in order to draw more robust conclusions. Also, practitioners in assessment of creativity in educational setting are encouraged to use more than one tool for a more comprehensive mastery of students' performance. For example, by doing a profile analysis of students' performance, educators can have a better understanding on students' strengths and weaknesses based on different measures, which might be helpful for a more differentiated instruction. Second, in testing MI, the present study gave some evidence of a psychometrically valid tool (with 15% to 50% problem items removed) for assessing DT as well as evaluative skill cross-culturally. The specific tasks that failed to meet MI point to the problem that although cross-cultural studies in creativity are emerging, it is not appropriate to compare observed means if the measurement invariance of the measures is not satisfied (Karwowski, 2016). Thus, researchers should be cautious in interpreting cross-cultural observed creativity scores in cases where MI is not tested, and future studies investigating MI issues in the field of creativity are needed to establish a set of measures that is appropriately invariant across many cultures.

Limitations

There are several limitations in this study. First, it concerns the measures. The present study only employed two items (subtests) for each of the four types of DT tests, and six ideas presented for each evaluation task. Future studies can investigate DT tests with three or more items and evaluation tasks with a larger number of ideas to see whether there are differences in the psychometric properties of the measures cross-culturally. Also, the ideas for evaluation based on Instances and Consequences tasks were from previous studies with a small to medium sample size, which may trivially affect the measurement of evaluative accuracy. Future studies may compile ideas from studies with a relatively large sample size, for example, from the DT tests of the present study. In addition, there are various types of DT tests in the field, and the present study only explored a subset of them. To make a more convincing conclusion regarding DT and evaluative skill, more types of DT tests as well as evaluation tasks should be included.

The second limitation concerns the sample. The participants in this study were college students in China and the US, thus the findings may not be generalized to populations of other age levels and educational backgrounds, or individuals in other countries. To establish a more convincing conclusion regarding differences between East Asians and Western populations, respondents of other age levels (e.g., children), educational backgrounds (e.g., individuals without college education) and from other countries (e.g., Japan and Finland) can be investigated. In addition, the females were a little overrepresented in the present study, and thus, a more balanced gender proportion is encouraged to be explored in the future.

Lastly, the constructs being studied can be broadened. For example, the examination of evaluative skill should go beyond evaluating others' ideas. Future studies can further investigate cross-cultural differences in situations where people are evaluating their own ideas. Additionally, the investigation of cultural orientation should not be limited to individualism-collectivism and uncertainty avoidance. Specifically, the dimensions of power distance and others which may be promising in predicting creativity can be further explored.

References

Acar, S., & Runco, M. A. (2019). Divergent thinking: New methods, recent research, and extended theory. *Psychology of Aesthetics, Creativity, and the Arts, 13*(2), 153–158. https://doi.org/10.1037/aca0000231

Acar, S., Runco, M. A., & Park, H. (2020). What should people be told when they take a divergent thinking test? A meta-analytic review of explicit instructions for divergent thinking. *Psychology of Aesthetics, Creativity, and the Arts, 14*(1), 39-49. https://doi.org/10.1037/aca0000256

Adair, W. L., & Xiong, T. X. (2018). How Chinese and Caucasian Canadians conceptualize creativity: the mediating role of uncertainty avoidance. *Journal of Cross-Cultural Psychology*, *49*(2), 223-238. https://doi.org/10.1177%2F0022022117713153

Averill, J. R., Chon, K. K., & Hahn, D. W. (2001). Emotions and creativity, east and west. *Asian Journal of Social Psychology*, *4*(3), 165-183. https://doi.org/10.1111/1467-839X.00084

Ball, O. E., & Torrance, E. P. (1978). Culture and tendencies to draw objects in internal visual perspective. *Perceptual and Motor Skills*, *47*(3, Pt 2), 1071-1075.

Barbot, B. (2019). Measuring creativity change and development. *Psychology of Aesthetics, Creativity, and the Arts, 13*, 203–210. https://doi.org/10.1037/aca0000232

Basadur, M. (1995). Optimal ideation-evaluation ratios. *Creativity Research Journal, 8*, 63–76.

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods, 22*(3), 507. https://psycnet.apa.org/doi/10.1037/met0000077

Belzak, W., & Bauer, D. J. (2020). Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning. *Psychological Methods, 25*(6), 673. https://psycnet.apa.org/doi/10.1037/met0000253

Benedek, M., Nordtvedt, N., Jauk, E., Koschmieder, C., Pretsch, J., Krammer, G., & Neubauer, A. C. (2016). Assessment of creativity evaluation skills: A psychometric investigation in prospective teachers. *Thinking Skills and Creativity, 21*, 75-84. https://doi.org/10.1016/j.tsc.2016.05.007

Brew, M. & Chen, Y. (2007) 'Where (who) are collectives in collectivism? Towards conceptual clarification on individualism and collectivism', *Psychological Review, 114* (1), 133–151.

Browne, M.W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K.A. Bollen & J.S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456. https://doi.org/10.1037/0033-2909.105.3.456.

Byrne, B. M., & Watkins, D. (2003). The issue of measurement invariance revisited. *Journal of Cross-cultural Psychology, 34*(2), 155-175. https://doi.org/10.1177/0022022102250225

Campbell, D. N. (1960). Blind variation and selective retention in creative thought as in other knowledge processes. *Psychological Bulletin, 67*, 380-400.

Charles, R. E., & Runco, M. A. (2001). Developmental Trends in the Evaluative and Divergent Thinking of Children. *Creativity Research Journal, 13*(3/4), 417-437. https://doi.org/10.1207/S15326934CRJ1334_19

Chaves, J. (2003). Confucianism: The conservatism of the East. *Intercollegiate Review*, *38*(2), 44.

Chen, F. F., & West, S. G. (2008). Measuring individualism and collectivism: The importance of considering differential components, reference groups, and measurement invariance. *Journal of Research in Personality*, *42*(2), 259-294. https://doi.org/10.1016/j.jrp.2007.05.006

Chen, G.-M., & Chung, J. (1994). The impact of Confucianism on organizational communication. *Communication Quarterly*, *42*(2), 93-105. https://doi.org/10.1080/01463379409369919

Chen, Z., Mo, L., and Honomichl, R. (2004). Having the memory of an elephant: long-term retrieval and the use of analogues in problem solving. *J. Exp. Psychol. Gen. 133*, 415–433. https://doi.org/10.1037/0096-3445.133.3.415

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

Cheung, P. C., Lau, S., Lubart, T., Chu, D. H. W., & Storme, M. (2016). Creative potential of Chinese children in Hong Kong and French children in Paris: A cross-cultural comparison of divergent and convergent-integrative thinking. *Thinking Skills and Creativity*, *22*, 201-211. https://doi.org/10.1016/j.tsc.2016.09.005

Christensen, P. R., Merrifield, P. R., & Guilford, J. P. (1953). *Consequences form A-1*. Beverly Hills, CA: Sheridan Supply.

Chua, R. Y., Roth, Y., and Lemoine, J. F. (2015). The impact of culture on creativity: how cultural tightness and cultural distance affect global innovation crowdsourcing work. *Adm. Sci. Q. 60*, 189–227. https://doi: 10.1177/0001839214563595

Cordon, S. L., & Finney, S. J. (2008). Measurement invariance of the Mindful Attention Awareness Scale across adult attachment style. *Measurement and Evaluation in Counseling and Development, 40*(4), 228-245. https://doi.org/10.1080/07481756.2008.11909817

Cramond, B., Kim, K. H., Chiang, T. W., Higuchi, T., Iwata, T., Ma, M., Palaniappan, A. K. (2020). Trends and challenges of creativity development among selected Asian countries and regions: China, Hong Kong/Macau, Japan, Malaysia and South Korea. In S. R. Smith (Ed.), *Handbook of giftedness and talent development in the Asia-Pacific* (pp. 1107-1133). Springer International Handbooks of Education. https://doi.org/10.1007/978-981-13-3021-6_51-1

Cropley, A. (2006). In praise of convergent thinking. *Creativity Research Journal, 18*(3), 391-404. https://doi.org/10.1207/s15326934crj1803_13

Deng, L., Wang, L., & Zhao, Y. (2016). How creativity was affected by environmental factors and individual characteristics: A cross-cultural comparison perspective. *Creativity Research Journal, 28*(3), 357-366. https://doi.org/10.1080/10400419.2016.1195615

Erez, M., & Nouri, R. (2010). Creativity: The influence of cultural, social, and work contexts. *Management and Organization Review*, *6*(3), 351-370.

Fee, A., & Gray, S. J. (2012). The expatriate-creativity hypothesis: A longitudinal field test. *Human Relations, 65* (12), 1515–1538. https://doi.org/10.1177/0018726712454900

Finch, W. H., & French, B. F. (2015). *Latent variable modeling with R*. New York, NY: Routledge.

Garcia-Barrera, M. A., Kamphaus, R. W., & Bandalos, D. (2011). Theoretical and statistical derivation of a screener for the behavioral assessment of executive functions in children. *Psychological Assessment, 23*(1), 64. https://psycnet.apa.org/doi/10.1037/a0021097

Grohman, M., Wodniecka, Z., & Klusak, M. (2006). Divergent Thinking and Evaluation Skills: Do They Always Go Together? *Journal of Creative Behavior, 40*(2), 125-145. https://doi.org/10.1002/j.2162-6057.2006.tb01269.x

Guilford, J. P. (1967). *The nature of human intelligence*. New York: McGraw-Hill.

Guilford, J. P. (1979). *Cognitive psychology with a frame of reference*. San Diego, CA: EDITS.

Guo, J., Ge, Y., & Pang, W. (2019). The underlying cognitive mechanisms of the rater effect in creativity assessment: The mediating role of perceived semantic distance. *Thinking Skills and Creativity*, *33*, 100572. https://doi.org/10.1016/j.tsc.2019.100572

Guo, J., & Guo., Y. (2021). *The influence of culture on creativity and its underlying motivational mechanisms*. Unpublished manuscript.

Guo, J., Lin, S., Guo, Y. (2018) Sex, birth order, and creativity in the context of China's one-child policy and son preference. *Creativity Research Journal*, *30*(4), 361-369. https://doi.org/10.1080/10400419.2018.1530535

Guo, Y., Lin, S., Guo, J., Lu, Z. L., & Shangguan, C. (2021). Cross-cultural measurement invariance of divergent thinking measures. *Thinking Skills and Creativity, 41*, 100852. https://doi.org/10.1016/j.tsc.2021.100852

Hahm, C. (2002). Law, culture, and the politics of Confucianism. *Columbia Journal of Asian Law*, *16(2)*, 254-301.

Hansen, M., Cai, L., Stucky, B. D., Tucker, J. S., Shadel, W. G., & Edelen, M. O. (2014). Methodology for developing and evaluating the PROMIS® smoking item banks. *Nicotine & Tobacco Research, 16* (Suppl_3), S175-S189. https://doi.org/10.1093/ntr/ntt123

Harpe, S. E. (2015). How to analyze Likert and other rating scale data. *Currents in Pharmacy Teaching and Learning, 7*(6), 836-850. https://doi.org/10.1016/j.cptl.2015.08.001

Hartley, K. A., Plucker, J. A., & Long, H. (2016). Creative self-efficacy and teacher ratings of student creativity in Chinese elementary classrooms. *Thinking Skills and Creativity, 22*, 142–151. https://doi.org/10.1016/j.tsc.2016.10.001

Hofstede, G. (1980a). *Culture's consequences: International differences in work-related values*. Beverly Hills, CA: Sage.

Hofstede, G. (1980b). Motivation, leadership, and organization: Do American theories apply abroad? *Organizational Dynamics, 16*(4), 42–63.

Hofstede, G. (1991). *Cultures and organizations: Software of the mind*. London, England: McGraw-Hill.

Hofstede, G., & Bond, M. H. (1984). Hofstede's culture dimensions: An independent validation using Rokeach's value survey. *Journal of Cross-cultural Psychology, 15*(4), 417-433.

Hoyle, R. H., & Panter, A. T. (1995). Writing about structural equation models. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 158–176). Thousand Oaks, CA: Sage.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55. http://doi.org/10.1080/10705519909540118

Ivancovsky, T., Shamay-Tsoory, S., Lee, J., Morio, H., & Kurman, J. (2019). A dual process model of generation and evaluation: A theoretical framework to examine cross-cultural differences in the creative process. *Personality and Individual Differences*, *139*, 60-68. https://doi.org/10.1016/j.paid.2018.11.012

Jellen, H. G., & Urban, K. K. (1989). Assessing creative potential world-wide: the first cross-cultural application of the test for creative thinking—drawing production (TCT-DP). *Gifted Education International*, *6*(2), 78-86.

Jones, G. K., & Davis, J. (2000). National culture and innovation: Implications for locating global R&D operations. *Management International Review, 40*, 1–39.

Jung, J. M. (2002). *The interactive impact of culture and individual characteristics on ethical decision-making processes, criteria, and judgmental outcomes: A cross-cultural comparison between South Korea and United States* (Doctoral dissertation, University of Cincinnati, Cincinnati, USA).

Karwowski, M. (2016). Culture and psychometric studies of creativity. In V. P. Glăveanu, (Ed.), *The Palgrave handbook of creativity and culture research* (pp. 159-186). Palgrave Macmillan.

Karwowski, M., Gralewski, J., Patston, T., Cropley, D. H., & Kaufman, J. C. (2020). The creative student in the eyes of a teacher: A cross-cultural study. *Thinking Skills and Creativity, 35,* Article 100636. https://doi.org/10.1016/j.tsc.2020.100636

Kasof, J. (1995). Explaining creativity: The attributional perspective. *Creativity Research Journal, 8*, 325–370. https://doi.org/10.1207/s15326934crj0804_1

Kim, D. (2016). *Role of cognitive flexibility in bilingualism and creativity* (Doctoral dissertation, University of Georgia, Athens, USA).

Kim, K. H. (2009). Cultural influence on creativity: The relationship between Asian culture (Confucianism) and creativity among Korean educators. *The Journal of Creative Behavior*, *43*(2), 73-93. https://doi.org/10.1002/j.2162-6057.2009.tb01307.x

Kim, K. H., Cramond, B., & Bandalos, D. L. (2006). The latent structure and measurement invariance of scores on the Torrance Tests of Creative Thinking–Figural. *Educational and Psychological Measurement*, *66*(3), 459-477. https://doi.org/10.1177/0013164405282456

Kim, K. H., Lee, H. E., Chae, K.-B., Anderson, L., & Laurence, C. (2011). Creativity and Confucianism Among American and Korean Educators. *Creativity Research Journal, 23*(4), 357. https://doi.org/10.1080/10400419.2011.621853

Kozbelt, A. (2007). A quantitative analysis of Beethoven as self-critic: Implications for psychological theories of musical creativity. *Psychology of Music, 35*(1), 144-168. https://doi.org/10.1177%2F0305735607068892

Krumm, G., Filipppetti, V. A., Lemos, V., Koval, J., & Balabanian, C. (2016). Construct validity and factorial invariance across sex of the Torrance Test of Creative Thinking–Figural Form A in Spanish-speaking children. *Thinking Skills and Creativity, 22*, 180-189. https://doi.org/10.1016/j.tsc.2016.10.003

Krumm, G., Lemos, V., & Filippetti, V. A. (2014). Factor structure of the Torrance Tests of Creative Thinking Figural Form B in Spanish-speaking children: Measurement invariance across gender. *Creativity Research Journal, 26*(1), 72-81. https://doi.org/10.1080/10400419.2013.843908

Kuhn, J. T., & Holling, H. (2009). Measurement invariance of divergent thinking across gender, age, and school forms. *European Journal of Psychological Assessment*, *25*(1), 1-7. https://doi.org/10.1027/1015-5759.25.1.1

Licuanan, B. F., Dailey, L. R., & Mumford, M. D. (2007). Idea evaluation: Error in evaluating highly original ideas. *The Journal of Creative Behavior, 41*(1), 1-27. https://doi.org/10.1002/j.2162-6057.2007.tb01279.x

Lonergan, D. C., Scott, G. M., & Mumford, M. D. (2004). Evaluative aspects of creative thought: Effects of appraisal and revision standards. *Creativity Research Journal*, *16*, 231-246. https://doi.org/10.1080/10400419.2004.9651455

Long, H., & Plucker, J. A. (2015). Assessing creative thinking: Practical applications. In R. Wegerif, L. Li, & J. C. Kaufman (Eds.). *The Routledge international handbook of research on teaching thinking* (pp. 315–329). Abingdon: Routledge.

Lucas, H. C., & Goh, J. M. (2009). Disruptive technology: How Kodak missed the digital photography revolution. *The Journal of Strategic Information Systems, 18*(1), 46-55. https://doi.org/10.1016/j.jsis.2009.01.002

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, *7*(4), 361-388. https://doi.org/10.1177/1094428104268027

Morris, M. W., & Leung, K. (2010). Creativity East and West: Perspectives and parallels. *Management and Organization Review*, *6*, 313-327.

Mueller, J., Melwani, S., Loewenstein, J., & Deal, J. J. (2018). Reframing the decision-makers' dilemma: Towards a social context model of creative idea recognition. *Academy of Management Journal, 61*(1), 94-110. https://doi.org/10.5465/amj.2013.0887

Mumford, M. D. (2001). Something old, something new: Revisiting Guilford's conception of creative problem solving. *Creativity Research Journal*, *13*, 267-276. https://doi.org/10.1207/S15326934CRJ1334_04

Mumford, M. D., Mobley, M. I., Uhlman, C. E., Reiter-Palmon, R., & Doares, L. (1991). Process analytic models of creative capacities. *Creativity Research Journal, 4*, 91–122.

Muthén, L. K. & Muthén, B. O. (1998-2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.

Ng, A. K. (2001). Why Asians are less creative than Westerners. Singapore, Singapore: Prentice Hall.

Ng, A. K. (2003). A cultural model of creative and conforming behavior. *Creativity Research Journal*, *15*(2-3), 223-233. https://doi.org/10.1080/10400419.2003.9651414

Ng, A. K., & Smith, I. (2004). The paradox of promoting creativity in the Asian classroom: An empirical investigation. *Genetic, Social, and General Psychology Monographs, 130*(4), 307–332. https://doi.org/10.3200/MONO.130.4.307-332

Nijstad, B. A., De Dreu, C. K., Rietzschel, E. F., and Baas, M. (2010). The dual pathway to creativity model: creative ideation as a function of flexibility and persistence. *European Review of Social Psychology, 21*(1), 34–77. http://doi.org/10.1080/ 10463281003765323

Niu, W., & Kaufman, J. C. (2013). Creativity of Chinese and American cultures: A synthetic analysis. *The Journal of Creative Behavior, 47*(1), 77–87. https://doi.org/10.1002/jocb.25

Niu, W., & Sternberg, R. J. (2002). Contemporary studies on the concept of creativity: The East and the West. *The Journal of Creative Behavior, 36*(4), 269-288. https://doi.org/10.1002/j.2162-6057.2002.tb01069.x

Niu, W., & Sternberg, R. J. (2006). The philosophical roots of Western and Eastern conceptions of creativity. *Journal of Theoretical and Philosophical Psychology, 26*(1-2), 18-38. https://psycnet.apa.org/doi/10.1037/h0091265

Okuda, S. M., Runco, M. A., & Berger, D. E. (1991). Creativity and the finding and solving of real-world problems. *Journal of Psychoeducational Assessment, 9*, 45-53.

Oyserman, D. (2006). High power, low power: and equality: culture beyond individualism and collectivism. *Journal of Consumer Psychology, 16*, 352–357. https://doi.org/10.1207/s15327663jcp1604_6

Paek, S. H., Abdulla, A., Acar, S., & Runco, M. A. (2021). Is more time better for divergent thinking? A meta-analysis of the time-on-task effect on divergent thinking. *Thinking Skills and Creativity*, 100894. https://doi.org/10.1016/j.tsc.2021.100894

Plucker, J. A. (1999). Is the proof in the pudding? Reanalyses of Torrance's (1958 to present) longitudinal data. *Creativity Research Journal, 12*(2), 103-114. https://doi.org/10.1207/s15326934crj1202_3

Plucker, J. A., Beghetto, R. A., & Dow, G. T. (2004). Why isn't creativity more important to educational psychologists? Potentials, pitfalls, and future directions in creativity research. *Educational Psychologist, 39*(2), 83-96. https://doi.org/10.1207/s15326985ep3902_1

Plucker, J. A., Qian, M., & Schmalensee, S. L. (2014). Is what you see what you really get? Comparison of scoring techniques in the assessment of real-world divergent thinking.

Creativity Research Journal, *26*(2), 135-143. http://doi.org/10.1080/10400419.2014.901023

Puente-Diaz, R., Maier, M. A., Brem, A., & Cavazos-Arroyo, J. (2016). Generalizability of the four C model of creativity: A cross-cultural examination of creative perception. *Psychology of Aesthetics, Creativity, and the Arts*, *10*(1), 14. https://psycnet.apa.org/doi/10.1037/aca0000038

Reiter-Palmon, R., Forthmann, B., & Barbot, B. (2019). Scoring divergent thinking tests: A review and systematic framework. *Psychology of Aesthetics, Creativity, and the Arts, 13*(2), 144. https://psycnet.apa.org/doi/10.1037/aca0000227

Reiter-Palmon, R., Wigert, B., & de Vreede, T. (2012). Team creativity and innovation: The effect of group composition, social processes, and cognition. In M. D. Mumford (Ed.), *Handbook of organizational creativity* (pp. 327–357). Waltham, MA: Elsevier.

Rensvold, R. B., & Cheung, G. (2001). Testing for metric invariance using structural equation models: Solving the standardization problem. In C. A. Schriesheim & L. L. Neider (Eds.), *Equivalence in measurement* (Vol. 1, pp. 25–50). Greenwich, CT: Information Age.

Rhodes, M. (1961). An analysis of creativity. *The Phi Delta Kappan, 42*(7), 305-310.

Rinne, T., Steel, G. D., & Fairweather J. (2012). Hofstede and Shane revisited: The role of power distance and individualism in national- level innovation success. *Journal of Cross-Cultural Research, 46*, 91–108. https://doi.org/10.1177/1069397111423898

Rinne, T., Steel, G. D., & Fairweather, J. (2013). The role of Hofstede's individualism in national-level creativity. *Creativity Research Journal*, *25*(1), 129-136. https://doi.org/10.1080/10400419.2013.752293

Rodgers, E. M., & Adhikarya, R. (1979). Diffusion of innovations: Up to date review and commentary. In D. Nimmo (Ed.), *Communications Yearbook 3* (pp. 67–81). New Brunswick, NJ: Transaction.

Runco, M. A. (1991). The evaluative, valuative, and divergent thinking of children. *The Journal of Creative Behavior, 25*(4), 311-319. https://doi.org/10.1002/j.2162-6057.1991.tb01143.x

Runco, M. A. (1999). Divergent thinking. In M. A. Runco, & S. R. Pritzker (Eds.). *Encyclopedia of creativity* (pp. 577-582). Academic Press.

Runco, M. A. (2004). Personal creativity and culture. In S. Lau, ANN. Hui, & GYC. Ng (Eds.), *Creativity: When east meets west* (pp. 9-21). World Scientific Publishing Company.

Runco, M. A. (2020). Idea evaluation. In M. A. Runco, & S. R. Pritzker (Eds.), *Encyclopedia of creativity* (3rd ed., pp. 607-611). Academic Press.

Runco, M. A., Abdulla, A. M., Paek, S. H., Al-Jasim, F. A., & Alsuwaidi, H. N. (2016). Which test of divergent thinking is best?. *Creativity. Theories–Research-Applications, 3*(1), 4-18. https://doi.org/10.1515/ctra-2016-0001

Runco, M. A., & Chand, I. (1994). Problem finding, evaluative thinking, and creativity. In M. A. Runco (Ed.), *Problem finding, problem solving, and creativity* (pp. 40–76). Norwood, NJ: Ablex.

Runco, M. A., & Chand, I. (1995). Cognition and creativity. *Educational Psychology Review, 7*, 243-267. https://doi.org/10.1007/BF02213373

Runco, M. A., & Dow, G. T. (2004). Assessing the accuracy of judgments of originality on three divergent thinking tests. *The International Journal of Thinking & Problem Solving, 14*(2), 5-14.

Runco, M. A., & Jaeger, G. J. (2012). The standard definition of creativity. *Creativity Research Journal*, *24*(1), 92-96. https://doi.org/10.1080/10400419.2012.650092

Runco, M. A., & Smith, W. R. (1992). Interpersonal and intrapersonal evaluations of creative ideas. *Personality and Individual Differences*, *13*(3), 295-302. https://doi.org/10.1016/0191-8869(92)90105-X

Runco, M. A., & Vega, L. (1990). Evaluating the creativity of children's ideas. *Journal of Social Behavior & Personality*, *5*(5), 439-452.

Rusticus, S. A., & Hubley, A. M. (2006). Measurement invariance of the Multidimensional Body-Self Relations Questionnaire: Can we compare across age and gender? *Sex Roles, 55*, 827–842. doi:10.1007/s11199-006-9135-7

Saeki, N., Fan, X., & Dusen, L. (2001). A comparative study of creative thinking of American and Japanese college students. *Journal of Creative Behavior, 35*(1), 24–36. https://doi.org/10.1002/j.2162-6057.2001.tb01219.x

Said-Metwaly, S., Fernández-Castilla, B., Kyndt, E., & Van den Noortgate, W. (2020a). Testing conditions and creative performance: Meta-analyses of the impact of time limits and instructions. *Psychology of Aesthetics, Creativity, and the Arts, 14*(1), 15–38. https://doi.org/10.1037/aca0000244

Said-Metwaly, S., Van den Noortgate, W., & Barbot, B. (2020b). Torrance test of creative thinking-verbal, Arabic version: Measurement invariance and latent mean differences across gender, year of study, and academic major. *Thinking Skills and Creativity*. https://doi.org/10.1016/j.tsc.2020.100768 (2020).

Shane, S. (1993). Cultural influences on national rates of innovation. *Journal of Business Venturing, 8*, 59–73.

Shao, Y., Zhang, C., Zhou, J., Gu, T., & Yuan, Y. (2019). How does culture shape creativity? A mini-review. *Frontiers in Psychology*, *10*, 1219. https://doi.org/10.3389/fpsyg.2019.01219

Shavitt, S., Lalwani, A. K., Zhang, J., & Torelli, C. J. (2006). The horizontal/vertical distinction in cross-cultural consumer research. *Journal of Consumer Psychology, 16*, 325–342. https://doi.org/10.1207/s15327663jcp1604_3

Silvia, P. J. (2008). Discernment and creativity: how well can people identify their most creative ideas? *Psychology of Aesthetics, Creativity and the Arts, 2*, 139–146. https://doi.org/10.1037/1931-3896.2.3.139

Simonton, D. K. (1988). *Scientific genius: A psychology of science*. New York: Cambridge University Press.

Simonton, D. K. (1999). Creativity as blind variation and selective retention: Is the creative process Darwinian? *Psychological Inquiry, 10*, 309-328.

Singelis, T. M., Triandis, H. C., Bhawuk, D. P., & Gelfand, M. J. (1995). Horizontal and vertical dimensions of individualism and collectivism: A theoretical and measurement refinement. *Cross-cultural Research*, *29*(3), 240-275.

Sternberg, R. J. & Lubart, T. I. (1995). *Defying the crowd*. New York: Free Press.

Sturm, A., Kuhfeld, M., Kasari, C., & McCracken, J. T. (2017). Development and validation of an item response theory ‐ based Social Responsiveness Scale short form. *Journal of Child Psychology and Psychiatry, 58*(9), 1053-1061. https://doi.org/10.1111/jcpp.12731

Sullivan, G. M., & Artino Jr, A. R. (2013). Analyzing and interpreting data from Likert-type scales. *Journal of Graduate Medical Education, 5*(4), 541-542. https://doi.org/10.4300/JGME-5-4-18

Tan, C. H. P. (2016). Creativity and Confucius. *Journal of Genius and Eminence, 1*(1), 79-84. https://doi.org/10.18536/jge.2016.01.1.1.10

Tang, M., Werner, C., & Karwowski, M. (2016). Differences in creative mindset between Germany and Poland: The mediating effect of individualism and collectivism. *Thinking Skills and Creativity, 21*, 31-40. https://doi.org/10.1016/j.tsc.2016.05.004

Thompson, M. S., & Green, S. B. (2006). Evaluating between-group differences in latent variable means. In G. R. Hancock, & G. R. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 119–169). Information Age Publishing.

Torrance, E. P. (1966). *The Torrance Tests of Creative Thinking—Norms, Technical Manual Research Edition—Verbal Tests, Forms A and B—Figural Tests, Forms A and B.* Personnel Press.

Torrance, E. P. (1974). *The Torrance Tests of Creative Thinking—Norms, Technical Manual Research Edition—Verbal Tests, Forms A and B—Figural Tests, Forms A and B.* Personnel Press.

Triandis, H. C. (1995). *Individualism and collectivism*. Boulder, CO: Westview.

Triandis, H. C., & Gelfand, M. J. (1998). Converging measurement of horizontal and vertical individualism and collectivism. *Journal of Personality and Social Psychology*, *74*(1), 118.

Wallach, M., & Kogan, N. (1965). *Modes of thinking in young children*. New York: Holt, Rinehart, & Winston.

Wallas, G. (1926). *The art of thought*. New York: Harcourt Brace.

Wicherts, J. M., Dolan, C. V., & Hessen, D. J. (2005). Stereotype threat and group differences in test performance: A question of measurement invariance. *Journal of Personality and Social Psychology, 89*, 696–716. https://doi.org/10.1037/0022-3514.89.5.696

Williams, Z. J., & Gotham, K. O. (2021). Improving the measurement of alexithymia in autistic adults: a psychometric investigation and refinement of the twenty-item Toronto Alexithymia Scale. *Molecular Autism*, *12*(1), 1-23. https://doi.org/10.1186/s13229-021-00463-5

Wonder, J., & Blake, J. (1992). Creativity east and west: Intuition vs. logic?. *The Journal of Creative Behavior*, *26*(3), 172-185. https://psycnet.apa.org/doi/10.1002/j.2162-6057.1992.tb01174.x

Wu, A. D., Zhen, L., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research, and Evaluation*, *12*(1), 3. https://doi.org/10.7275/mhqa-cd89

Xie, G., & Paik, Y. (2019). Cultural differences in creativity and innovation: are Asian employees truly less creative than western employees? *Asia Pac. Bus. Rev. 25*, 123–147. https://doi.org/10.1080/13602381.2018.1535380

Xu, L., & Barnes, L. L. (2011). Measurement invariance of scores from the Inventory of School Motivation across Chinese and US college students. *International Journal of Testing, 11*(2), 178–210. https://doi.org/10.1080/15305058.2010.542357

Yao, X., Wang, S., Dang, J., & Wang, L. (2012). The role of individualism-collectivism in the individual creative process. *Creativity Research Journal, 24*(4), 296-303. https://doi.org/10.1080/10400419.2012.730001

Yuan, K. H., & Bentler, P. M. (2006). Mean comparison: Manifest variable versus latent variable. *Psychometrika, 71*(1), 139–159. https://doi.org/10.1007/s11336- 004-1181-x

Zha, P., Walczyk, J. J., Griffith-Ross, D. A., Tobacyk, J. J., & Walczyk, D. F. (2006). The impact of culture and individualism–collectivism on the creative potential and achievement of American and Chinese adults. *Creativity Research Journal*, *18*(3), 355-366. https://doi.org/10.1207/s15326934crj1803_10

Zhao, Q. (2016). *A test of group incubation in creativity study field and the exploration of its underlying mechanisms*. (Master's Thesis, East China Normal University, Shanghai, China).

Appendix A

Divergent Thinking Tests-Line Meanings

Directions: On the next few sections, you are going a play a game.
You will see some lines and figures and after you have looked at each one, please tell us all of the things that the drawing makes you think of.
List all the things that the drawing could be. Think of as many things as you can. The more, the better. You can look at it from any direction (or angle) you want.

Appendix B

Divergent Thinking Tests-Uses

Directions: On the next page, you will be asked to list all the uses you can think of for some daily objects.
Please list as many uses as you can for the following objects. The more, the better.

What uses can you think of for a **chair**?

What uses can you think of for a **rope**?

Appendix C

Divergent Thinking Tests-Instances

Directions: On the next page, you will be asked to do another creativity task. For this task, please list as many instances as possible of things with the following features. The more, the better.

What instances can you think of things that are **square**?

What instances can you think of things that can **move on wheels**?

Appendix D

Divergent Thinking Tests-Consequences

Directions: On the next page, you will be asked to do another creativity task. For this task, please list as many consequences as possible if the following situation happened. The more, the better.

What would happen as a consequence if people **lost the ability of reading and writing**?

What would happen as a consequence if **people no longer needed food in life**?

Appendix E

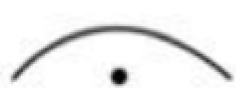
Evaluative Measures-Ideas from Line Meanings

Now that you got some experience on performing those interesting games, the following are some ideas produced by others. You will have the opportunity to evaluate their ideas. There is no right or wrong answer.

For each of the three figures listed below, people were asked to produce as many ideas as possible about what the drawing could be. They can look at it from any direction (or angle) they want.

Below are the ideas produced by them, and you are going to evaluate the ideas regarding how many people out of 10 would think of the idea. Small number means it is rare for people to think of this idea, and large number means many people can think of this idea.

Game 1:



	1	2	3	4	5	6	7	8	9	10
wifi										
mountain										
music										
sun										
golf ball										
breast										

Game 2

	1	2	3	4	5	6	7	8	9	10
camel										
worm										
string										
mountain										
signature										
river										

Appendix F

Evaluative Measures-Ideas from Uses

For each of the three daily objects listed below, people were asked to produce as many ideas as possible about what the uses could be.
Below are the ideas produced by them, and you are going to evaluate the ideas regarding how many people out of 10 would think of the idea. Small number means it is rare for people to think of this idea, and large number means many people can think of this idea.

Task 1: List all the uses for an **umbrella**.
Please evaluate each idea regarding how many people out of 10 would think of the use.

	1	2	3	4	5	6	7	8	9	10
collect water										
lampshield										
walking stick										
decoration										
weapon										
golf club										

Task 2: List all the uses for a **book**.
Please evaluate each idea regarding how many people out of 10 would think of the use.

	1	2	3	4	5	6	7	8	9	10
weapon										
conversation starter										
art										
to hold love letters										
to make a fire										
doorstop										

Appendix G

Evaluative Measures-Ideas from Instances

For each of the two tasks listed below, people were asked to produce as many ideas as possible about instances of things with specific features.
Below are the ideas produced by them, and you should evaluate the ideas regarding how many people out of 10 would think of the idea. Small number means it is rare for people to think of this idea, and large number means many people can think of this idea.

Task 1: List all the things that are **round**.
Please evaluate each idea regarding how many people out of 10 would think of the instance.

	1	2	3	4	5	6	7	8	9	10
neutron										
sink										
end of a ball pen										
fingertip										
earth										
doorknob										

Task 2: List all the things that can make a **noise**.
Please evaluate each idea regarding how many people out of 10 would think of the instance.

	1	2	3	4	5	6	7	8	9	10
thunder										
air conditioner										
walking										
stomach										
bees										
paper										

Appendix H

Evaluative Measures-Ideas from Consequences

For each of the two imagined situations listed below, people were asked to produce as many ideas as possible about consequences following that situation.
Below are the ideas produced by them, and you should evaluate the ideas regarding how many people out of 10 would think of the idea. Small number means it is rare for people to think of this idea, and large number means many people can think of this idea.

Situation 1: What would happen as a consequence **if people no longer needed to sleep**?
Please evaluate each idea regarding how many people out of 10 would think of the consequence.

	1	2	3	4	5	6	7	8	9	10
not happy										
no dreams										
no need for a bed										
more crime										
party longer										
more fights										

Situation 2: What would happen as a consequence **if people could live twice longer than now**?
Please evaluate each idea regarding how many people out of 10 would think of the consequence.

	1	2	3	4	5	6	7	8	9	10
technological advances										
more divorces										
be wiser										
more sicknesses										
more wrinkles										
food shortages										

Appendix I

The Individualism-Collectivism Test

Directions: Please use a scale from 1 = disagree (false) to 9 = agree (true) to indicate your agreement or disagreement with the following statements.

1. I'd rather depend on myself than others.
2. It is important that I do my job better than others.
3. If a coworker gets a prize, I would feel proud.
4. Parents and children must stay together as much as possible.
5. I rely on myself most of the time; I rarely rely on others.
6. Winning is everything.
7. The well-being of my coworkers is important to me.
8. It is my duty to take care of my family, even when I have to sacrifice what I want.
9. I often do "my own thing."
10. Competition is the law of nature.
11. To me, pleasure is spending time with others.
12. Family members should stick together, no matter what sacrifices are required.
13. My personal identity, independent of others, is very important to me.
14. When another person does better than I do, I get tense and aroused.
15. I feel good when I cooperate with others.
16. It is important to me that I respect the decisions made by my group.

Appendix J

The Uncertainty Avoidance Test

Direction: Please use a scale from 1 = strongly disagree (false) to 7 = strongly agree (true) to indicate your agreement or disagreement with the following statements.

1. I tend to get anxisou easily when I don’t know an outcome.
Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree
2. I feel stressful when I cannot predict consequences.
Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree
3. I dislike unpredictable situations.
Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree
4. I dislike it when a person’s statement could mean different things.
Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree
5. I don’t like to go into a situation without knowing what I can expect from it.
Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree
6. I don’t like situations that are uncertain.
Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Appendix K

Demographic Measures

What is your gender?
Male
Female
I'm not sure
Other

How old are you? (please indicate a number)

What is your major?

What ethnicity do you identify yourself with?
White
Latino
Black or African American
Asian
American Indian or Alaska Native
Native Hawaiian or Pacific Islander
Multiracial or Other