A comparative evaluation of seven instruments for measuring values comprising Hofstede's model of culture

By: Vas Taras, Piers Steel, Madelynn Stackhouse

Taras, V., Steel, P., & Stackhouse, M. (2023). A comparative evaluation of seven instruments for measuring values comprising hofstede's model of culture. Journal of World Business, 58(1). DOI: 10.1016/j.jwb.2022.101386

Abstract:

Culture and its measurement are foundational to International Business research. Hofstede's model of culture dominates cross-cultural research. Unfortunately, the evidence of poor psychometric properties of Hofstede's instrument for measuring cultural values, the VSM, has been mounting, which prompted the development of numerous alternative instruments for measuring cultural values comprising Hofstede's model of culture. The abundance of choices makes it challenging to determine which of the instruments is most suitable for a given study. Using a large international sample (N = 12,462), we evaluated the psychometric properties of seven different instruments for measuring individual-level values in Hofstede's cultural framework and assessed their content validity, reliability, factor structure, and measurement equivalence. Our tests confirmed that Hofstede's instrument suffers from several psychometric deficiencies, while other instruments, notably those developed by Dorfman and Howell (1988), Yoo et al. (2011), and Taras et al. (2013), showed good reliability and validity. Guidelines for selecting the most suitable instrument and directions for future instrument development are provided.

Keywords: cultural values | Hofstede | measurement of values | instrument comparison

Article:

1. Introduction

In his classic book "Culture's Consequences," Hofstede (1980) introduced a model for understanding culture from an "objective" measurement perspective that could meaningfully be used to understand and explain differences across national cultures. Since Hofstede (1980) introduced his framework, research on culture in international business has surged, with the number of publications increasing yearly (cf., Taras et al., 2012). Indeed, according to Google Scholar, the different editions of Hofstede's book "Culture's Consequences" have been cited over 100,000 times, and "Culture and Organization: Software of the Mind" has gathered tens of thousands of citations (Hofstede, 1980, 2001; Hofstede et al., 2005, 2010), making Hofstede by far the most cited scholar in international business research and cross-cultural studies.

Importantly, Hofstede's model of culture was accompanied by the Value Survey Module (VSM), a data collection instrument for quantifying cultural values, allowing researchers to not only use the national cultural indices published in "Culture's Consequences" but also collect their own data describing the cultural values of their research subjects. Although several other instruments for measuring culture were developed before Hofstede (e.g., Kluckhohn & Strodtbeck, 1961; Rokeach, 1973) and many more after the release of VSM (Taras et al., 2009), VSM has remained the instrument of choice in the majority of studies that required measuring cultural values of study participants (cf., meta-analytic reviews by Taras et al., 2012, 2010).

Despite its popularity, the VSM has been criticized for its numerous deficiencies, notably for its poor psychometric properties, the use of confusing constants and items weights in scoring, and problems with its factor structure and measurement invariance (Baskerville, 2003; Bearden et al., 2006; Gerlach & Eriksson, 2021; Kruger & Roodt, 2007; McSweeney, 2002; Spector et al., 2001). Additionally, Hofstede repeatedly warned that his model and instrument were designed exclusively for the national level of analysis. The enormous popularity of Hofstede's cultural framework, coupled with the ever-wider recognized limitations of his VSM instrument, prompted many researchers to develop their own alternatives to VSM. In fact, over a hundred instruments for measuring cultural values are available, many specifically designed to measure cultural values comprising Hoftede's model (Taras et al., 2009).

Given the persistent need to measure culture and the overwhelming choice of instruments, it is surprising that there has never been a comprehensive empirical comparison of the psychometric properties of the instruments available for the task. Indeed, prior publications have focused either on evaluating Hofstede's VSM (e.g., Kruger & Roodt, 2007; Spector et al., 2001) or introducing new instruments (e.g., Maznevski & DiStefano, 1995; Yoo et al., 2011), each study based on a different sample and conducted under different conditions rendering a direct comparison of these reports unreliable. While these studies provide important insights into the underpinnings and validity of each measure, they do not offer an empirical comparison of the psychometric properties and the strengths and weaknesses of each measure relative to one another.

The present study seeks to fill this gap by conceptually and empirically comparing several popular cultural value instruments designed for measuring Hofstede's cultural values at the individual level. The quality of the data determines the quality of the study's findings, and the choice of the data collection instrument is perhaps the most important decision a researcher makes (cf. Schriesheim et al., 2001). In more metaphorical terms, if we do not consider measurement issues, "we may wind up erecting theoretical skyscrapers on foundations of empirical jello" (Schriesheim et al., 2001, p. 516).

Our comparison of the instruments included the following assessment. First, using a panel of 158 subject matter experts (i.e., faculty in international business positions), we critically examine each instrument's face and content validity. Second, we recruited an international group of 12,462 participants who were randomly assigned to complete one of the seven instruments and used these data to evaluate and compare the internal consistency of each instrument. Third, using the same international participant data, we assessed the instruments' factor structures using exploratory and confirmatory factor analyses. Fourth, we tested the measurement equivalence of the scales for gender, student status, previous job experience, and whether participants spoke English fluently or came from an English-speaking home country. Fifth, we examined the impact of using different scales by meta-analytically determining their collective method variance across eight variables. Finally, based on our comparative analysis, we offered guidance for researchers

seeking to either select the most suitable scale from among existing instruments for measuring cultural values or for developing new instruments.

2. Measuring culture: a historical overview

The measurement of culture, defined as a set of relatively stable values, norms, beliefs, traditions, and artifacts that are shared in a given population (cf., the review of definitions in Taras et al. (2009), began in the field of anthropology, which relied primarily on descriptive qualitative research methods such as ethnographic research. However, as Lord Kelvin put it in 1883, "When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind" (Thomson, 1883). Without the ability to measure culture quantitatively, business researchers showed little interest in culture. All of that changed in 1980 when Hofstede published "Culture's Consequences" and a new measure for measuring culture from a question-and-answer format not subject to rater differences (the 16-item Values Survey Module). When conducting an employee survey commissioned by IBM in 1967–73, Hofstede found systemic differences in how employees from different countries responded to the VSM. By aggregating the responses to the VSM questions in each country sample, Hofstede produced national cultural rankings that opened countless new opportunities in international business and management research. Attempts to measure culture by asking the respondents to answer a standard set of standard questions had been undertaken before Hofstede (e.g., Kluckhohn & Strodtbeck, 1961; Rokeach, 1973). However, Hofstede was the first to conduct a study based on a large international sample.

Despite its enormous popularity, Hofstede's work has been the subject of criticism. First, there have been concerns with Hofstede's conceptualization of culture, with some scholars claiming that "Hofstede never studied culture" (Baskerville, 2003, p. 1) and that the popularity of his work is based on "a triumph of faith [and] a failure of analysis" (McSweeney, 2002, p. 89). As a result, the practice of using Hofstede's national cultural rankings published in "Culture's Consequences" as a proxy for culture used in hundreds of studies (Taras & Steel, 2009) has been widely criticized. Second, Hofstede's national rankings were derived based on the data from 1967 to 73, raising questions about their validity decades later in the light of well-documented cultural changes (e.g., Inglehart & Baker, 2000; Taras et al., 2012). Third, there are questions about the generalizability of the IBM sample to the general populations of the countries where the survey was conducted (McSweeney, 2002). As noted by Schwartz (1994), "a well-educated and well-compensated IBM employee was probably a good representation of the general population in industrialized countries such as the United States or Germany, but far from the general population in developing countries such as El Salvador or Pakistan" (p. 91). Fourth, the VSM instrument used to collect the data has been shown to have problems with low reliability, unstable factor structure, and measurement invariance (Bearden et al., 2006; Gerlach & Eriksson, 2021; Kruger & Roodt, 2007; Spector et al., 2001). Perhaps in response to this criticism, and to his credit, Hofstede never stopped refining his model and instrument (cf. Hofstede et al., 2010).

In sum, despite the many concerns, Hofstede's framework has largely defined how we conceptualize and measure culture. For example, Taras et al. (2009) reviewed 121 instruments for measuring cultural values and concluded that "97.5% of all reviewed measures contain at least some dimensions that are conceptually similar to those introduced by Hofstede" (p. 360). Further,

self-report questionnaires have become the pervasive approach to studying culture in international business (Taras & Steel, 2009).

3. Hofstede's cultural value framework

Hofstede's culture model originally included four values or cultural dimensions: Individualism-Collectivism, Power Distance, Masculinity-Femininity, and Uncertainty Avoidance. Later, Long-Term Orientation was added to the dimensions and became an integral part of the model (Hofstede & Bond, 1988). As mentioned, Indulgence vs. Restraint was recently added to the model in response to the World Values Survey (Hofstede et al., 2010). However, perhaps because of its recency, the Indulgence vs. Restraint dimension has not reached the same level of popularity across measurement instruments, and thus, we omit it from our review.

Individualism-Collectivism describes the relationship between the individual and the group. People with individualist values put their interests above those of the group and believe everyone is expected to look after themselves and their immediate family (Hofstede, 1980, 2001). Conversely, collectivists put the group's interests first based on the belief that people from birth onward are integrated into a strong, cohesive in-group.

Individualism-Collectivism is the most popular dimension in cross-cultural research. As noted by Taras et al. (2014), "cross-cultural research of the past decades has been, to a large degree, research on individualism" (p. 214). Indeed, reviews of cross-cultural research comment on the ubiquity of Individualism in the literature (Kirkman et al., 2006; Tsui et al., 2007), and meta-analyses of studies on the effects of cultural values in the workplace showed that over 85% of all empirical evidence related to culture is about Individualism (Taras et al., 2010, 2012).

It is worthwhile noting that the Individualism cultural dimension "took on a life of its own" and evolved in the literature beyond Hofstede's original definition (Taras et al., 2009; House et al., 2004). Under the umbrella of Individualism-Collectivism, we can often see preferences as a preference to work in teams versus work alone, the attitudes toward work-life balance, attitudes towards the family, the roles and relationships among the family members, and more. To a large extent, Individualism has become a "catch-all" cultural concept for the West, as has Collectivism for the East. This has led to confusion about the definitions and boundaries of the construct. Because of this, some subsequent models created multiple dimensions to disentangle these distinct components. For example, the GLOBE model differentiates between Institutional and In-Group Collectivism (House et al., 2004). There also have been attempts to distinguish between the vertical and horizontal types of Individualism and Collectivism (Singelis et al., 1995). In Hofstede's view, however, Individualism and Collectivism are the opposites of the same continuum (Hofstede, 1980, 1994, 2001; Hofstede et al., 2010).

Power Distance describes the attitudes toward people in power. It is "the extent to which the less powerful persons in a society [or organization] accept inequality in power and consider it as normal" (Hofstede, 1986, p. 307). People with high Power Distance values believe and accept that supervisors and subordinates in organizations, or more and less senior family members, have fundamentally different rights and obligations and should enjoy different privileges and respect. In contrast, people with low Power Distance values do not display the same reverence and respect for seniority.

There are two issues related to Power Distance that are worthwhile noting. First, much research, including from Hofstede himself, reported a strong correlation between Individualism and Power Distance, and some even questioned if Individualism and Power Distance are different

dimensions (cf., Minkov, 2018; Minkov & Kaasa, 2020). While there may be a strong empirical connection between the two dimensions, perhaps to the point of functional equivalence, conceptually the two dimensions describe different types of values and tendencies. Second, the GLOBE study (House et al., 2004) has shown that the same individual may report a different Power Distance orientation depending on how the question is framed. Specifically, when asked what kind of "boss" you want to be versus what kind of "boss" you want to have, people in some cultures, notably the U.S., report very different preferences (e.g., want to be authoritative, suggesting high Power Distance values, but want to have a democratic supervisor suggesting low Power Distance values).

*Masculinity-Femininit*y is defined as "the degree to which such masculine values as advancement, earnings, training, up-to-datedness [versus] such feminine values as the friendly atmosphere, position security, physical conditions, and cooperation are valued" (Hofstede, 2001, p. 281). Masculine cultures stress material success and assertiveness; feminine cultures stress the quality of life, interpersonal relationships, and concerns for the weak or, in other words, this is "the degree to which values like assertiveness, performance, success, and competition […] prevail over values like the quality of life, maintaining warm personal relationships, service, care for the weak, and solidarity" (Hofstede, 1994, p. 6).

As noted by Minkov & Kaasa (2020), Masculinity-Femininity is among the most "disjointed" dimensions in Hofstede's model. First, as defined above, Hofstede states that Masculinity-Femininity is about valuing achievement and advancement versus group harmony and warm relationships. However, by invoking the concept of "masculinity" and "femininity," Hofstede tied this dimension to gender roles and norms. Furthermore, he noted that in masculine societies, gender roles are, indeed, more differentiated (i.e., men are expected to be more assertive and aggressive, whereas women are expected to be kind and caring), while, in contrast, such gender role differentiation is less pronounced in feminine societies (Hofstede, 2001). As a result of this confusion, Masculinity-Femininity is often discussed in the literature in the context of gender roles and gender role differentiation, which conceptually is very different from values of achievement orientation (e.g., Lippa, 2001). To avoid confusion, the GLOBE model (House et al., 2004) splits Masculinity into the need for Achievement Orientation versus Gender Egalitarianism, and, indeed, their data show that the two are distinct dimensions.

Uncertainty Avoidance describes the preference for clear rules and guidance. Specifically, it is "the extent to which people within a culture are made nervous by situations which they perceive as unstructured, unclear, or unpredictable, situations which they, therefore, try to avoid by maintaining strict codes of behavior and a belief in absolute truth" (Hofstede, 1986, p. 308). One confusion that arises around the definition of Uncertainty Avoidance is the notion of 'preference for rules' versus 'risk-aversion.' As Hofstede (2001) stresses, Uncertainty Avoidance "should not be confused with risk-avoidance" (p. 145).

Long-Term Orientation is perhaps the dimension with the least conceptual clarity. As noted by Minkov et al. (2018), confusion around the construct may have arisen from its naming ambiguity. Originally it was introduced under the name "Confucian Dynamism" (Bond & Hofstede, 1989; Hofstede & Bond, 1988). For people unfamiliar with Confucius's teachings, the term "Confucian Dynamism" is not very telling, and the word "dynamism" does not add clarity. However, even those who have studied Confucius extensively may remain puzzled, as his teachings cover a broad range of issues (Minkov et al., 2018).

In light of the conceptual confusion around the construct with its initial label, Hofstede re-labeled the construct as Long-Term Orientation. Hofstede & Bond (1988) originally stated that

Confucian Dynamism "deals with the choice from Confucius' ideas, and its positive pole reflects a dynamic, future-oriented mentality, whereas its negative pole reflects a more static, tradition-oriented mentality" (p. 16). The re-labeled dimension was defined as "the values on the one pole oriented towards the future (especially perseverance and thrift) and on the opposite pole more oriented towards the past and present" (Hofstede, 1991, p. 348). In his review of this dimension, Smith (2002) notes that Hofstede "first considers the nature of what was earlier named as Confucian work dynamics. Since both poles of the dimension appear to reflect Confucian values, he chooses, perilously, to focus instead primarily upon the single item that loaded most heavily on the positive pole ('persistence') and rename the dimension as Long-Term Orientation" (p. 131). As such, in its current form, Long-Term Orientation refers to the values of persistence and thrift, personal stability, and respect for tradition. In contrast, Short-Term Orientation refers to the tendency to live in the moment, enjoy the now, and give little consideration to the future.

4. The instruments evaluated in this study

For our comparative assessment, we selected English language versions of the cultural value instruments that met the following three inclusion criteria. First, they must evaluate at least the four original dimensions in Hofstede's model, preferably all five. Second, they must use Hofstede's approach to data collection, namely the self-report questionnaire. Third, the instruments must have been designed with the expressed purpose of measuring cultural values at the individual level of analysis. Out of almost 150 instruments in Taras' (2019) Catalogue of Instruments for Measuring culture, seven instruments met these criteria, which includes Hofstede's VSM instrument. Although Hofstede repeatedly warned against its use at the individual level of analysis (Hofstede, 2002, 2006), it has been so used in hundreds of studies (cf., Taras et al., 2010, 2012), which warrants its investigation and comparison with the other instruments in this study.

4.4.1. Dorfman & Howell (1988)

Dorfman & Howell (1988) developed an instrument that rivals Hofstede's VSM in popularity. As of this writing, the book chapter that introduced the instrument (Dorfman & Howell, 1988) has been cited over 1500 times. The instrument is designed to measure the four original dimensions from Hofstede's model. Long-Term Orientation was added to Hofstede's model later, explaining its absence from Dorfman and Howell's instrument. In their publication, Dorfman & Howell (1988) criticized the quality of Hofstede's VSM, citing it as a reason for developing their own instrument. Based on their assessment, the internal reliability statistics (Cronbach's alphas) were generally acceptable (above 0.70), but in a few cases, were as low as 0.57, depending on the dimension and the sample. Later assessments by Culpepper & Watts (1999) also showed support for the factor structure of Dorfman and Howell's instrument. Notably, the published version of the instrument also contained the Paternalism dimension, defined as the belief that managers should "take a personal interest in worker's lives, provide for workers' personal needs, and generally take care of the workers" (Dorfman & Howell, 1988, p. 131). This dimension does not directly correspond to any of the values in Hofstede's model and has therefore been frequently dropped by researchers using the instrument. Indeed, the shorter four-dimensional version of the instrument has been commonly used in the literature.

4.4.2. Furrer et al. (2000)

Furrer et al. (2000) developed their instrument to mirror Hofstede's entire five-factor model. It is a concise instrument with exactly four items per dimension, each phrased as a value statement. The authors provided little information about the instrument's psychometric properties, but noted that it was an attempt to improve Hofstede's VSM. Even though the instrument was introduced almost "in passing," it gained considerable popularity among cross-cultural researchers. As of this writing, the paper was cited over 1000 times, and most citations mention the publication specifically in reference to the cultural value instrument, often by the scholars who used it in later studies.

4.4.3. Taras et al. (2013)

The IWORC (Individual Work-Related Culture) instrument was developed by Taras et al. (2013) to provide a more fine-grained picture of the cultural values that comprise Hofstede's framework. One way the instrument tried to address the deficiencies of Hofstede's VSM was to remove the confusion caused by lumping distinct values into one dimension, as happens with Individualism and Masculinity in Hofstede's model. Accordingly, IWORC splits the construct into Individualism as in personal interests versus the group's interests, which is how the dimension was originally defined by Hofstede, and the Preference to Work in Teams versus Work Alone. Also, as per Minkov's (2018) and Minkov & Kaasa's (2020) analyses, Masculinity proved to be a disorderly dimension. To recover coherence, consistent with the GLOBE's approach, IWORC splits the Masculinity dimension into Achievement Orientation and Gender Egalitarianism.

4.4.4. Vitell et al. (2003)

Vitell et al. (2003) developed their instrument to include all five values in Hofstede's model. Notably, Vitell et al. (2003) refer to the fifth dimension as Confucian Dynamism and not as Long-Term Orientation. The paper that introduced the instrument provided the results of an assessment of the instrument's quality and noted that they sometimes were substandard. For example, the reported Cronbach's alphas were in the 0.61 to 0.67 range for Power Distance, Masculinity, and Individualism, although they exceeded the 0.70 thresholds for the other two dimensions.

4.4.5. Voich (1995)

Voich (1995) is the lengthiest instrument in our collection, containing 31 items in total. However, in addition to measuring Hofstede's four original dimensions, it also measures Paternalism and Work Ethics. Both dimensions do not directly correspond to any of the dimensions in Hofstede's model. It defines Paternalism as the belief that managers must be fatherly figures to their employees, provide for the needs of their subordinates, and take care of them, and it defines Work Ethic as the belief that work is good in itself and that it will bestow dignity on a person who has a high level of work ethic" (p. 34). Although neither dimension completely corresponds to any of the five dimensions in Hofstede's model, in this study, we used the complete six-dimensional version of the instrument, as other scholars usually use it.

4.4.6. Yoo et al. (2011)

Yoo et al. (2011) CVSCALE (Cultural Values Scale) was explicitly developed to "measure Hofstede's five dimensions of cultural values at the individual level" (p. 193). It is growing steadily in popularity, approaching 200 citations annually. The instrument's psychometric properties were assessed in four different samples, and the instrument was found to meet or exceed the quality standards generally, although some Cronbach's alphas were between 0.62 and 0.68, depending on the sample and the dimension. Yoo et al. (2011) also provided the results of exploratory factor analysis and showed that all factor loadings and goodness of fit indices met expectations. In a follow-up study, Yoo & Shin (2017)) reviewed the reported reliability estimates from a broad range of studies, finding average results from 0.72 to 0.78.

4.4.7. Hofstede's vsm (VSM-94)

Lastly, we also assessed Hofstede's VSM. There are several different versions of the instrument. Originally, VSM-82 was introduced to the public. However, the instrument had several glaring limitations. For example, it combined Likert-type and multiple-choice scales (selecting the preferred type of manager), raising questions about the compatibility and appropriateness of using these drastically different types of items to derive composite scores. Also, some questions referred to specific conditions at IBM (e.g., "Are employees afraid to disagree with their managers?"), which seems more a measure of IBM's organizational culture than the respondent's values. The VSM-94 version addressed some of these issues and is the one most commonly used by other researchers; accordingly, we used this version in our study.

Notably, all versions of VSM retained the peculiar scoring that involved constants and item weights. For example, to calculate the total Individualism score, the following formula was used: IND = −27(mean A6) + 30(mean A8) + 76(mean A12) – 43(A18) −29

The reason for this complex scoring scheme is that Hofstede wanted to emulate T scores, so he used constants and weightings to force the results into a 0 to 100 range. Despite several adjustments, this strategy worked only in part. The final results reported in "Culture's Consequences" (2001) included negative values and values above 100. Such weightings were arbitrary and problematic, and Hofstede kept recalibrating them, trying to keep the scores within the 0–100 range. Unsurprisingly, when VSM was subsequently used by other scholars to measure individual values, it became a standard practice to use a Likert-type 5-point scale and simply average the responses, as it is done with the other instruments reviewed here. Importantly, the present study does not aim to compare national averages but to assess the instrument's content validity, internal consistency, and hypothesized factor structure. These assessments require raw unweighted data without the scale totals or the item weights and constants.

5. Method

5.1. Participants and data

A sample of 12,462 respondents was used to collect the data for the comparative assessment of the seven instruments. The participants were MBA and undergraduate business students from over 200 universities in 65 countries (165 countries by the country of origin – born and raised - of students) who participated in the X-Culture international business competition (Taras & Ordeñana, 2015).

For the list of countries as well as specific demographics for each of them, see Table S1 in the Online Supplement*. Working in teams of 6–7, the project participants serve as international business consultants for client companies worldwide and typically conduct market research, identify new business opportunities, and develop market entry strategies for different countries.

The average age of the respondents was 23.1 years, ranging from 16 to 67. A little over 85.5% of the sample had at least some work experience, averaging 3.6 years and ranging from 0 to 20+ years. At the time of the survey, 58.1% had jobs, of which 63.9% involved the supervision of at least one person and 36.1% did not involve supervision. All respondents were conversational in English, confirmed by a short English test, self-report, and post-project peer evaluations. The average TOEFL (Test of English as a Foreign Language) score was 9.1 out of 10. Accordingly, all data were collected in English.

Participants were randomly assigned to complete one of the seven instruments. On average, 1376 respondents fully completed each instrument. The participants also reported their age, gender, work experience, and nationality, which we used for assessing measure invariance.

All measures were completed on a 5-point scale (e.g., strongly disagree to strongly agree; not important to very important; always to never). The complete list of items for each scale is provided in the Online Supplement*.

To compare the quality of the seven instruments, we proceeded as follows. First, we evaluated the content validity of the instruments. Next, we assessed the reliability/ internal consistency of each dimension (Cronbach's alpha). Third, we proceeded with the exploratory factor analysis to test each scales' factor structure stability. Fourth, we conducted confirmatory factor analyses to test the construct validity and measurement equivalence. Finally, we meta-analytically determined the random-effects variance component (REVC), which in this context is almost entirely the method variance attributable to the different scales. With it, we can establish credibility intervals, which is how much correlations can shift due to scale selection. We generated correlations based on the Big Five Personality traits (John & Srivastava, 1999), Cultural Intelligence (Ang et al., 2007), Emotional Intelligence (Wong & Law, 2002), and Self-Efficacy (adapted version of Sherer et al., 1982).

6. Results

To compare the quality of the seven instruments, we drew upon the modern requirements for scale development, as per Boateng et al. (2018) and Kyriazos & Stalikas (2018). Missing data were handled with listwise deletion specific to each analysis.

6.1. Content validity

First, we evaluated the degree to which the content of each instrument reflects the value dimensions in Hofstede's conceptual model. We used the Content Validity Index (CVI), as suggested by Polit & Beck (2006). Using the Academy of International Business directory, itself an international group with members from 72 countries, we invited academics specializing in cross-cultural and international business research to evaluate each survey and rate it on its suitability to assess their corresponding contracts. Following the CVI methodology, the 158 experts were presented with the full collection of items from all seven instruments associated with one of the five dimensions in Hofstede's model, accompanied by a definition of the dimension the items purported to represent, and subsequently rated each item according to its suitability/relevancy to the dimension (scored

from 1=Completely irrelevant/unsuitable to 5=Very relevant/suitable). Scores above 3 (i.e., Somewhat relevant/suitable) were deemed content valid. A link to the full copy of the survey, including instructions, is provided in the endnotes.

The total item count was: 49 Individualism, 36 Power Distance, 38 Masculinity, 35 Uncertainty Avoidance, and 32 Long-Term Orientation/Confucian Dynamics. Each item was evaluated by 31 to 36 experts depending on the dimension. To balance the number of items per expert (each expert was asked to evaluate approximately 50 items), dimensions with fewer items were blocked together (e.g., Confucian Dynamism and Masculinity, Uncertainty Avoidance and Teamwork, Achievement Orientation and Power Distance). As can be seen, we far exceed the thresholds recommended by Koo & Li (2016), who suggested that "researchers should try to obtain at least 30 heterogeneous samples and involve at least three raters whenever possible when conducting a reliability study" (p. 158) and the number of experts for content validation usually recommended as 5 to 8.

We used the intraclass correlation coefficient (ICC) to assess whether raters were able to follow instructions and recognize the content validity of the items. Given the raters are considered a sample of the larger pool of cultural researchers and each rated every item in their section, we used ICC(2) with absolute agreement. As per Koo & Li (2016), a score above 0.75 is considered good, and above 0.90 is considered excellent. In our case: Teamwork and Uncertainty Avoidance's ICC(2,30) = 0.89, Achievement and Power Distance's ICC(2,36) = 0.91, Confucian Dynamism and Masculinity ICC(2,26) = 0.92, and LTO and Individualism's ICC(2,21) = 0.93, confirming that, in aggregate, the experts largely concur with each other's ratings, justifying content validation.

Table 1 provides the results of the content validity analysis, highlighting two versions of the Content Validity Index (i.e., A and B). As can be seen, most of the items in the majority of the instruments appeared to relate to the constructs they were intended to measure. However, there were notable exceptions. Overall, the most content-valid was Vitell et al.'s scale, with Taras et al.'s a close second. In contrast, most items in Hofstede's VSM were judged by experts to be errant. For example, Individualism neared the floor for relevance suitability, with an average of 1.7 and only 25% of raters considering its items relevant. This poor showing by the VSM proved to be a common occurrence in subsequent analyses.

6.2. Scale reliability (Cronbach's alphas)

Table 2 provides the results of the internal consistency assessment. Cronbach's alpha is very sensitive to k (the number of items): the more items, the higher the coefficient. A good instrument is short yet provides a valid and reliable assessment of the construct in question. There is always a tradeoff between parsimony and internal consistency: when two instruments have similar Cronbach's alphas, the one with fewer items is usually preferred.

Perhaps not surprisingly, given the results of the content validity assessment presented earlier, Hofstede's VSM has the lowest Cronbach's alphas, at only 0.43 on average across all five dimensions, ranging from 0.18 for Power Distance and never exceeding the commonly accepted 0.70 threshold, though approaching it for Individualism and Long-Term Orientation. Similarly, none of the dimensions in the Furrer et al.'s instrument reached the 0.70 standard. In contrast, every dimension in the instruments by Dorfman and Howell, Vittel et al., Taras et al., and Yoo et al. exceeded the 0.70 standard, with the best showing by Taras et al. (average alpha 0.82) and Dorfman and Howell (average alpha 0.81). Notably, the scale by Taras et al. (2013) had the lowest number of items per dimension, only 3.57 on average, and the highest individual item reliability.

Table 1. Content validity.

Instruments	Dorfman & Howell		Furrer et al.		Vittel et al.		Voich		Taras et al.'s IWORC		Yoo et al.'s CVSCALE		Hofstede's VSM	
Dimensions	4		5		5		6		7		5		5	
Items	26		20		21		31		25		25		20	
	A	B	A	B	A	B	A	B	A	B	A	B	A	B
Individualism	4.3	86.9	3.5	71.3	4.5	95.1	4.1	82.4	4.3	87.0	4.2	82.3	1.7	25.0
Masculinity	2.6	46.3	3.5	76.7	3.8	82.2	2.6	44.1			2.7	52.4	3.3	69.3
Power Distance	3.5	76.5	3.6	75.0	3.8	81.5	3.5	75.9	3.9	83.9	3.4	73.1	3.0	61.2
Uncert. Avoid.	3.3	70.6	2.5	41.4	3.9	83.8	3.2	66.3	3.6	80.5	3.4	74.1	2.0	28.5
Long-term Or.			3.0	59.6					4.4	94.4	3.6	73.2	2.9	61.1
Confucian Dyn.					3.6	77.2								
Achievement									3.9	86.3				
Gender Egalit.									2.6	44.4				
Team Preference									4.3	88.5				
Overall	3.4	70.1	3.2	64.8	3.9	84.0	3.4	67.2	3.9	80.7	3.5	71.0	2.6	49.0
Rank	4		6		1		5		2		3		7	

Note. A and B represent two versions of the Content Validity Index. A: Average relevance/suitability score across items on a 1 to 5 scale. B: Percentage of experts on average who rated items as being 3 or higher on a 1 to 5 relevance/suitability scale.

Table 2. Internal consistency (Cronbach's Alphas).

Instrument	Dorfman & Howell		Furrer et al.		Vittel et al.		Voich		Taras et al.'s IWORC		Yoo et al.'s CVSCALE		Hofstede's VSM	
Dimensions	4		5		5		6		7		5		5	
Items	26		20		21		31		25		25		20	
Sample Size	1281		1291		1260		1242		3590		4131		968	
Individualism	6	0.72	4	0.54	3	0.71	4	0.61	4	0.81	6	0.72	4	0.68
Masculinity	9	0.89	4	0.43	4	0.73	5	0.85			4	0.76	4	0.23
Power Distance	6	0.76	4	0.48	5	0.74	5	0.73	4	0.76	5	0.83	4	0.18
Uncert. Avoid.	5	0.86	4	0.47	5	0.77	5	0.47	4	0.86	5	0.80	4	0.35
Long-Term Or.			4	0.54					3	0.77	6	0.78	4	0.69
Confucian Dyn.					4	0.78								
Paternalism							5	0.61						
Work Ethic							7	0.72						
Achievement									3	0.81				
Gender Egalitar.									4	0.87				
Team preference									3	0.84				
Average *k*	6.50		4.00		4.20		5.17		3.57		5.20		4.00	
Average Item Alpha		0.42		0.20		0.42		0.30		0.56		0.41		0.19
Average Alpha		0.81		0.49		0.74		0.67		0.82		0.78		0.43
Rank	2		6		4		5		1		3		7	

6.3. Exploratory factor analysis (EFA)

Next, using EFA, we assessed how well the items within each instrument relate to their respective dimensions (latent factors). Each instrument was subjected to an EFA using maximum likelihood extraction. We used a varimax rotation instead of an oblique rotation to be consistent with the underlying model that each dimension is orthogonal. The following criteria were used to evaluate the instrument quality.

The number of factors: First, we checked how many factors emerge based on the unrestricted EFA, relying on the scree plot and the number of Eigenvalues > 1.0. The analysis was then repeated with the number of factors limited to the number of dimensions specified for the instrument, and these were the results used to assess the factor loadings. If the number matched the number of dimensions in the underlying model, we considered it supportive. Conversely, if the instrument is based on a five-dimensional model, for example, but the EFA solution suggests that there are four or fewer latent factors, we take that as a problem with the instrument.

Loadings on respective latent factors: Next, using the rotated factor solution, we calculated what percentage of items loaded on their respective factors. We used 0.40 as the threshold for the loadings, consistent with Brown (2006). Substantively lower loadings suggest that the item may be measuring something else and does not belong in this group.

Misloadings and cross-loadings: Furthermore, all instruments and their underlying culture models assume that the cultural values included in the model are conceptually distinct and mostly empirically independent from one another (hence varimax rotation). Some relationships among the dimensions are possible, but overall, they are expected to represent different constructs. Accordingly, we wanted to see items load on one latent factor only, specifically the factor representing the dimension for that item. Items that load on multiple factors or items that load on a wrong factor indicate a problem with the instrument.

Table 3 presents the results of EFA. Consistently with prior analyses, Hofstede's VSM yet again showed subpar results. First, the solution suggested that there are two instead of five factors. When adopting a five-factor solution, over half of the items (i.e., eleven) either misloaded, failed to load, or cross-loaded. Finally, the average factor loading was 0.38, below the 0.40 cut-off. Consequently, the factor structure produced by EFA differed greatly from the underlying model.

Also problematic were Furrer et al.'s and Voich's instruments, both with eight items loading errantly. In contrast, the EFA produced factor structures that were very close to the hypothesized for the instruments by Dorfman and Howell, Vitell et al., Taras et al., and Yoo et al., all with the correct number of factors, the majority of item loading appropriately (i.e., between 80% and 96%), and large absolute factor loadings (i.e., between 0.59 and 0.70). The Taras et al. scale was ranked first, cleanly measuring seven factors with just 25 items, only one of which loaded errantly.

Table 3. Exploratory factor analysis.

	Dorfman & Howell	Furrer et al.	Vittel et al.	Voich	Taras et al.'s IWORC	Yoo et al.'s CVSCALE	Hofstede's VSM
N	1281	1291	1260	1242	3590	4131	968
Factors	4	5	5	6	7	5	5
Factors with variance > 1	4	5	5	8	7	5	6
Inflection point	4	5	5	5	7	5	2
Average absolute loading	0.62	0.47	0.61	0.50	0.70	0.59	0.38
Item Number	26	20	21	31	25	25	20
Normally Distributed	26	20	17	30	22	25	13
Items loading errantly	2	8	1	8	1	5	11
% load as expected	92.31	60.00	95.24	74.19	96.00	80.00	45.00
Rank	3	6	2	4	1	5	7

Note. Items loading errantly are misloading or cross-loading on another factor at 0.40 (or above) or failing to load at 0.40 (or above). Items with either skew or kurtosis above ± 2 are considered non-normally distributed.

Table 4. Confirmatory factor analysis.

	Dorfman & Howell	Furrer et al.	Vitell et al.	Voich	Taras et al.'s IWORC	Yoo et al.'s CVSCALE	Hofstede's VSM
N	1281	1291	1260	1242	3590	4131	968
X2	1436.63	1502.78	1419.69	3116.48	2410.58	2706.231	756.46
df	293	160	179	419	254	289	160
CFI	0.907	0.646	0.845	0.742	0.950	0.901	0.842
TLI	0.897	0.580	0.818	0.714	0.941	0.889	0.812
RMSEA	0.057	0.082	0.076	0.075	0.049	0.054	0.062
SRMR	0.070	0.092	0.085	0.104	0.060	0.067	0.060
Rank	2	7	5	6	1	3	4

Note. CFI = Comparative Fit Index; TLI = Tucker Lewis Index; RMSEA = Root Mean Square Error of Approximation; SRMR = Standardized Root Mean Square Residual.

6.4. Confirmatory factor analysis (CFA)

Next, using lavaan's (version 0.6–8) "cfa" function, we conducted a CFA, where we tested how well each of the seven models fits the data when the relationships between the items and their latent factors are specified in advance. Kline (2015) suggests reporting the model X2, the Comparative Fit Index (CFI), the Root Mean Square Error of Approximation (RMSEA), and the Standardized Root Mean Square Residual (SRMR). To this end, we also include the Tucker Lewis Index (TLI), also known as the Non-Normed Fit Index (NNFI). As per West, Taylor, and Wu (2012), the cut-off for good fit is: CFI > 0.95, RMSEA 〈 0.06, SRMR < 0.08, and TLI ≥ 0.95. Notably, the CFI and the TLI also support 〉 0.90 as an indicator of good fit (Fan et al., 1999).

Table 4 summarizes the goodness of fit indices for each instrument. Based on RMSEA and the SRMR, three scales fit well: Dorfman and Howell's, Taras et al.'s, and Yoo et al.'s instruments. Surprisingly, given its previous problems, Hofstede's VSM also showed acceptable results, just slightly exceeding the 0.06 cut-off for the RMSEA and meeting the SRMR cut-off. However, if we expand the fit indices to the CFI and TLI, only the IWORC approaches the >0.95 cut-off, though the instruments by Dorfman and Howell and Yoo et al. approach the lower alternative cut-off of 0.90. Furrer et al., Vitell et al., and Voich show uniformly bad fit. Notably, if we constrained Voich's instrument to just the dimensions in common with Hofstede, dropping Paternalism and Work Ethic, fit declined substantially more, with the RMSEA and SRMR falling to 0.09 and 0.14, respectively.

6.5. Measurement invariance

Lastly, we considered measurement equivalency among the instruments, testing whether they perform similarly or differently based on sample demographics. We conducted tests for measurement invariance, which is typically done by comparing three increasingly restrictive models (Hirschfeld & Von Brachel, 2014). First, we consider configural invariance, where each group is expected to have the same factor structure. Second, we test for weak or metric invariance, where factor loadings are constrained to be identical across groups. Third, we test for strong or scale invariances, in which both intercepts and factor loadings are constrained to be identical. Using lavaan's (version 0.6–8) "cfa" again, we specified group constraints and collected the appropriate comparisons across the three models using its "lavTestLRT" function. If the configural or baseline model does not indicate a good fit, indicating constructs do not hold true across groups, it precludes acceptance of a later and more constrained model. Particularly important is to attend to the ΔCFI, which should stay smaller than 0.01 for the model comparisons to be considered equivalent (Cheung & Rensvold, 2002).

We contrasted four groups when testing for measurement invariance: gender (males versus females), student status (undergraduates versus MBAs), employment experience (none versus some), and English language fluency. The results of the gender measurement invariance tests are provided in Table S2 (see Online Supplement*). Consistent with our previous CFAs, the instruments by Furrer et al., Vitell et al., and Voich did not produce an acceptable baseline model. Again, Dorfman and Howell, as well as Yoo et al., showed mixed results that are borderline for CFI and TLI but a good fit for RMSEA and SRMR. Taras et al.'s scale had the best fitting baseline model, with all indices exceeding threshold requirements. Finally, the VMS failed to converge in baseline, which was expected given that almost half of its items were not normally distributed. However, given that the weak invariance model is minimally different from the baseline for all

other measures, we used it as a proxy. In this case, we rank the VSM fourth, just below Yoo et al., but with the insufficient fit. The remaining three measures – Dorfman and Howell, Taras et al., and Yoo et al. – showed a ΔCFI well below 0.01 for weak and strong invariance, indicating that men and women likely interpret the scales identically and provide comparable results.

Next, we considered business undergraduates versus MBAs and employment experience versus none. The use of student samples has been debated for decades (e.g., Bernstein et al. 1975). Student samples are neither inherently good nor bad but depend partially on whether the focus is on internal validity (i.e., establishing causal relationships) or external validity and generalizability (Stevens, 2011). For the latter, it is primarily an empirical matter, where some relationships generalize, and others do not. Previous research by Hanel & Vione (2016) on cultural values and student populations indicates that students' responses were as variable as the general population, provided comparable psychometrics, with typically negligible differences, reaching a peak with a 0.21 mean standardized difference for personal sexual behavior, equivalent to a correlation 0.10 or a small effect size. Similarly, in an editorial on the matter of international business research (Bello et al., 2009), they report multiple examples where MBAs provide ecologically valid results for cultural research, though the recommendation is that results be corroborated routinely with a managerial or employee group.

The measurement equivalence results for undergraduates versus MBAs are reported in Table S3 (see Online Supplement*). Again, only Dorfman and Howell, Taras et al., and Yoo et al. exceed or closely approach our established cut-off scores for the baseline. In all cases, the subsequent ΔCFI was well below 0.01 for weak and strong invariance, indicating that the two are effectively interchangeable. Again, the best fitting scale was the Taras et al. scale. Table S4 (see Online Supplement*) extends this test by determining if there is a difference between those with work experience with those yet to be employed. Due to data gathering limitations, Yoo et al. did not have the associated information. Partly due to the reduced sample size, only one of the remaining measures showed adequate fit, the Taras et al. Importantly, it indicated both weak and strong invariance, confirming that business school students with and without work experience were interchangeable. Surprisingly, the VSM was below the cut-off of 0.08 for the SRMR. As per Table S4 for the VSM, those who have work experience and those who do not generate equivalent factor loadings (i.e., weak equivalence), but their intercepts or means will likely differ (i.e., not strong equivalence).

We also assessed the effect of English language fluency (Tables S5, S6, see Online Supplement*). Table S5 uses results from the TOEFL (Test of English as a Foreign Language), splitting the results at whether respondents obtained a perfect score of 10 (N = 5117) or 9 and below (N = 6189). The average score for the '9 and below' group was 8.44. Due to data availability, the VSM was based on the four-dimension model, dropping Long-Term Orientation. Table S6 (see Online Supplement) uses self-reported home country to classify respondents according to whether the primary language was English (N = 4100) or Other (N = 7329) in their home country. For both tables, once again, only Dorfman and Howell, Taras et al., and Yoo et al. exceed or closely approach our established cut-off scores for the baseline. For these scales, minor English language deficits do not appear to be of concern, with strong equivalence across the different subgroups.

Finally, we assessed the equivalence of the measures meta-analytically by examining the correlations that each scale's major dimensions generate with the Big Five personality traits (John & Srivastava, 1999), Cultural Intelligence (Ang et al., 2007), Emotional Intelligence (Wong & Law, 2002), and Self-Efficacy (adapted version of Sherer et al., 1982). If the scales are interchangeable, then the variation among them should be largely due to sampling error, with a

Table 5. Correlations between cultural instruments and eight individual difference variables.

	A	C	E	N	O	CI	EI	SE
Dorfman & Howell	(352)	(352)	(352)	(352)	(352)	(288)	(288)	(266)
Individualism	.18	.08	.04	−0.13	.11	.23	.02	.18
Masculinity	−0.19	−0.16	−0.11	.07	−0.01	−0.23	−0.21	−0.04
Power Distance	−0.16	−0.12	−0.06	.09	.05	−0.33	−0.21	−0.09
Uncert. Avoid.	.20	.22	−0.04	−0.04	.00	.13	.14	.16
Furrer et al.	(367)	(367)	(367)	(367)	(367)	(308)	(308)	(246)
Individualism	−0.09	−0.06	.01	.01	−0.01	.10	.01	.10
Masculinity	.06	−0.00	.02	.05	−0.02	−0.05	.02	.12
Power Distance	−0.22	−0.15	−0.05	.07	−0.14	−0.05	.02	−0.06
Uncert. Avoid.	−0.22	−0.14	−0.05	.07	−0.14	.12	.05	−0.06
Long-Term Or.	.24	.06	.11	−0.11	.07	.10	.01	.15
Vitell et al.	(325)	(325)	(325)	(325)	(325)	(300)	(300)	(253)
Individualism	.25	.13	.12	−0.15	.15	.27	.07	.37
Masculinity	.11	.21	.16	−0.11	.17	.17	.10	.22
Power Distance	−0.02	−0.04	−0.05	.01	−0.14	−0.02	−0.11	−0.05
Uncert. Avoid.	.19	.14	−0.05	.13	−0.06	.06	.03	.29
Confucian Dyn.	.33	.23	.05	−0.02	.16	.20	.10	.32
Voich	(313)	(313)	(313)	(313)	(313)	(293)	(293)	(266)
Individualism	−0.12	−0.02	.05	.02	.06	−0.02	.06	.11
Masculinity	−0.24	−0.15	−0.09	.10	−0.02	−0.04	.05	−0.03
Power Distance	−0.20	−0.09	−0.09	.05	−0.04	−0.07	.01	.03
Uncert. Avoid.	−0.05	.01	.04	.04	−0.04	.10	.08	.08

Table 5. Correlations between cultural instruments and eight individual difference variables. *(continued)*

	A	C	E	N	O	CI	EI	SE
Taras et al.'s IWORC	(2503)	(2503)	(2503)	(2503)	(2503)	(349)	(3119)	(285)
Individualism	0.15	0.08	0.05	−0.13	0.12	.12	.14	.09
Power Distance	−0.12	−0.09	0.00	0.05	−0.05	.02	.05	−0.01
Uncert. Avoid.	0.22	0.19	0.06	0.00	0.06	.04	.08	.09
Long-Term Or.	0.17	0.20	0.04	−0.08	0.13	.08	.08	.07
Achievement	−0.16	−0.05	0.05	−0.03	0.07	.04	.18	.07
Gender Inegalit.	−0.22	−0.21	−0.09	0.04	−0.06	−0.04	−0.03	−0.10
Team Preference	0.19	0.05	0.17	−0.13	0.18	.14	.14	.18
Yoo et al.'s CVSCALE	(2257)	(2257)	(2257)	(2257)	(2257)	(2721)	(2506)	(1992)
Individualism	0.21	0.13	0.08	−0.15	0.14	.22	.17	.23
Masculinity	−0.13	−0.13	−0.10	0.00	−0.07	−0.03	−0.02	−0.02
Power Distance	−0.26	−0.22	−0.09	0.04	−0.09	−0.08	−0.07	−0.04
Uncert. Avoid.	0.31	0.30	0.06	−0.08	0.06	.25	.22	.21
Long-Term Or.	0.28	0.38	0.14	−0.16	0.22	.34	.36	.25
Hofstede's VSM						(361)	(362)	(267)
Individualism	NA	NA	NA	NA	NA	.21	.27	.25
Masculinity	NA	NA	NA	NA	NA	.19	.27	.26
Power Distance	NA	NA	NA	NA	NA	.15	.16	.25
Uncert. Avoid	NA	NA	NA	NA	NA	−0.03	−0.07	.03
Cronbach's Alpha	.77	.78	.82	.80	.74	.91	.88	.91

Note. A = Agreeableness; C = Conscientiousness; E = Extraversion; N = Neuroticism; O = Openness to Experience; CI = Cultural Intelligence; EI = Emotional Intelligence; SE = Self-Efficacy. Sample size in parentheses for the correlates below.

resulting small tau and I2 (both measures of the residual heterogeneity among correlations after taking into account sampling error). Essentially, the residual heterogeneity represents method variance attributable to how each scale operationalizes its cultural dimensions. We ran the analyses in R (version 4.1.1) with metafor (version 3.0–2), based on restricted maximum likelihood estimation (i.e., method = "REML") and unbiased sampling error (i.e., vtype = "UB"). Data was available except for the VSM and the Big Five personality traits. Aggregating results across all variables (i.e., eight meta-analyses per cultural dimension), we obtained the average I2, average tau, and the 90% credibility interval or range (i.e., how much correlations can expect to vary depending on the scale). For Individualism, I2=65.2%, tau=0.07, and the 90% correlation range is 0.22. For Masculinity, I2=89.5%, tau=0.12, and the 90% correlation range is 0.39. For Power Distance, I2=66.2%, tau=0.07, and the 90% correlation range is 0.23. For Uncertainty Avoidance, I2=81.54%, tau=0.10, and the 90% correlation range is 0.33. This is considerable variation, especially for Masculinity and Uncertainty Avoidance. As reflected in Table 5, this means that results can completely shift direction, from the positive to the negative, depending on what scale was used.

7. Discussion

Hofstede's model of culture continues to dominate research, and his values of Individualism, Power Distance, Uncertainty Avoidance, Masculinity, and Long-Term Orientation remain standard in cross-cultural studies. Likewise, Hofstede's VSM remains a popular choice for measuring these values, including at the individual level of analysis, at it has been done in hundreds of studies (cf., Taras et al., 2010, 2012), despite Hofstede's repeated assertions that VSM was designed for the national level of analysis. Unfortunately, numerous earlier studies evaluating the quality of VSM at both individual and national levels of analysis confirmed its poor psychometric properties (e.g., Kruger & Roodt, 2007; Kuchinke & Ardichvili, 2001; Spector et al., 2001). The focus of the present study was specifically on the instruments for the individual level of analysis. As per Hofstede's warnings and the findings from prior evaluations of VSM, we conclude that VSM is not the best choice for researchers who conduct individual-level cultural research, and they need to search for alternatives.

Recognizing the limitations of VSM, over a hundred other instruments for measuring culture have been developed since the publication of Hofstede's (1980) "Culture's Consequences," many of them specifically for measuring Hofstede's cultural values. The abundance of options and the absence of a comparative evaluation of the instruments make choosing the proper instrument difficult. The contribution of the present study is that it provides an in-depth comparative analysis of six popular instruments plus the VSM and evaluates their suitability for measuring Hofstede's cultural values at the individual level of analysis. Several of them, notably those by Dorman & Howell (1988), Taras et al. (2013), and Yoo et al. (2011), showed good content validity, internal consistency, factor structure, and measurement equivalence. The Taras et al.'s instrument had the added benefit of being very brief (only 3.57 items per dimension) and differentiating between the different facets of Individualism (self-interest versus group interest and preference to work alone versus in a team) and Masculinity (achievement orientation and gender roles), similar to the approach used in the GLOBE model (House et al., 2004).

7.1. Guidelines for scale selection

Critically, our meta-analyses (Table 5) underscore how important scale selection can be. Our results may only sporadically replicate across scales and samples, with a study using one scale capable of generating radically different results from an otherwise identical study that uses another. Indeed, Schriesheim et al. (2001) assessment that we may have "foundations of empirical jello" (p. 516) appears most justified. In broader terms, this is an instance of the jingle-jangle fallacy in that we have somewhat different constructs going by the same name (i.e., jingle). However, we are not alone. Flake et al. (2017) review how that lack of validity evidence for scales is endemic across the personality and social psychology field. They also provide some pointed recommendations for scale selection.

Flake et al. (2017) view scale validation as a sequential three-phase process, with each scale having to pass the previous phases to be considered for the next. Phase 1, "Substantive," deals with construct conceptualization and content validity. Phase 2, "Structural," deals with reliability, factor analysis, and measurement equivalence. Phase 3, "External," deals with convergent, discriminant, and predictive validity. Phase 3 is considered an ongoing process, perhaps best addressed through dozens of studies and summarized meta-analytically, though predicated on first passing through Phase 1 and 2. The present study focuses on the first two phases. We found that the instruments by Dorfman & Howell (1988), Yoo et al. (2011), and the Taras et al. (2013) scale showed good to excellent results in almost every test. Despite that the remaining scales had problems overall, individual subscales of the other measures may still be promising, suggesting a mix-and-match strategy could be pursued. For example, Table 1 indicates that Vitell et al. (2003) Masculinity dimension was considered content valid, and Table 2 shows adequate reliability. Perhaps this single scale could be extracted if a dedicated Masculinity scale is strongly desired. Furthermore, if researchers have idiosyncratic construct definitions, they might choose a scale that matches their research goals.

As for Phase 3, we have the three contenders of Dorfman & Howell (1988), Yoo et al. (2011), and Taras et al. (2013), who made it through Phase 1 and 2. Also, some of the individual subscales from the other four measures might also be considered. Until a formal meta-analysis is conducted that compares these scales' correlates with a variety of criteria, it relies on the individual researcher to spot-check associations. As a final consideration, if researchers need to employ non-English language versions, scales already with several validated translations may be the deciding factor (e.g., CVScale).

7.2. Limitations and directions for future research

While this study compared the quality of a set of instruments developed for measuring culture per Hofstede's framework, we do not definitively determine "the winners and losers." No doubt, researchers will continue developing new instruments, and "the best" today may not compare favorably to its competition in the nearest future. Accordingly, the goal of this study was to highlight the already-existing abundance of choices and to provide a roadmap for such comparative evaluation; to review a set of basic yet effective assessments any researcher can use to either select the most suitable data collection instrument from among existing options or to guide the development of new scales. We encourage that scale and model development, and evaluation of their quality should continue.

To these ends, this study's strengths are its large international sample and comprehensive evaluation of the psychometric properties of several cultural values measurement instruments. Still, while the respondents in our sample came from 174 countries, almost a third were from the U.S., which can limit the generalizability of our findings. Likewise, the multi-national nature of our sample and the resulting linguistic diversity and possible cultural biases could affect how the respondents answered the survey questions, posing a threat to the validity of our result.

While psychometric assessments of cultural value measures have been conducted on numerous occasions in prior research, generally, it was with one instrument evaluated in isolation, making direct comparisons impossible. However, the results of this study should be considered within its constraints. First, as noted by Boateng et al. (2018), there are multiple steps towards developing and validating scales, of which typically only a subset of these steps was applied to older scales but are now routinely expected from newer developments. Effectively, these older scales are "grandfathered" into common usage, and their potentially superior replacements languish as they struggle to meet the ever-growing list of scale construction requirements. The result is less than ideal, as the field reverts to "established" scales despite their potentially undesirable psychometric properties. This study may be viewed as retroactive scale development, where we reconsider and contrast the psychometric properties of scales to put the culture assessment field on a firmer foundation. Consistent with the growing list of scales and scale development specifications, we could have considered newer individual scales, such as Sharma (2010) or Minkov et al. (2018) Individualism and Long-Term Orientation scales. At the least, we can confirm that these are indeed reasonable options to consider.

Second, keeping within Hofstede's framework does not preclude expanding or refining it. While we focused on English language versions of the scales, translations may demonstrate different psychometric properties. Further, as reviewed here, many scales have added additional dimensions or have decomposed dimensions, which warrants further investigation.

Third, instead of administrating the scales individually, as we did here, they can be administered en masse, then subjecting the resulting item list to factor analysis. The advantage of this approach is the ability to pick and choose the best scales and items from our previous collective efforts, a strategy used effectively in other topics (e.g., Steel, 2010).

Fourth, in the current study, we provided information on measurement invariance, but continued work should be done in this context. Though most of the respondents in this study had at least some work experience, and many were in their 30 s, 40 s, or even 50 s, there will still be uncertainty regarding the generalizability of our findings and conclusions to samples of more mature organizational employees. In general, there is a belief that student populations are de facto, not equivalent to work samples, though the justification for this segregation is sparse. Age and education are at times weakly associated with culture, but typically far less than gender, generation, occupation, and socioeconomic status (Steel & Taras, 2010; Taras et al., 2016). Given so, it is extremely unlikely that university students represent a monolithic cultural group, but likely reliable differences depend on program or degree (i.e., occupation) and tuition (e.g., socioeconomic status). Our results here indicate an internationally diverse group of business students can be largely interchangeable with MBAs, with or without work experience, who themselves are often equivalent to a general working sample (Bello et al., 2009). Rather than automatically rejecting student samples, we should take a more nuanced approach and establish the boundary conditions of their generalizability, a long, long overdue task.

Finally, we assessed the scales' content validity, internal consistency (Cronbach's alphas), factor structure (EFA and CFA) of the instruments, and measurement equivalence. Given the array

of scales, we were not able to give each an exhaustive psychometric analysis (e.g., determining coefficients of stability) or provide full construct validation (i.e., discriminant, convergent, predictive). Particularly predictive validity appears relevant and important in the context of choosing the most suitable instrument for measuring culture. Some of the information needed for such assessment is already in the literature and needs meta-analytic review to consolidate it. Where it is lacking, we encourage future researchers to include these assessments in their studies to take the discussion presented here to the next level.

References

Ang, S., Van Dyne, L., Koh, C., Ng, K. Y., Templer, K. J., Tay, C., & Chandrasekar, N. A. (2007). Cultural intelligence: Its measurement and effects on cultural judgment and decision making, cultural adaptation and task performance. Management and Organization Review, 3(3), 335–371.

Baskerville, R. F. (2003). Hofstede never studied culture. Accounting, Organizations and Society, 28(1), 1–14.

Bearden, W. O., Money, R. B., & Nevins, J. L. (2006). Multidimensional versus unidimensional measures in assessing national culture values: The Hofstede VSM 94 example. Journal of Business Research, 59(2), 195–203.

Bello, D., Leung, K., Radebaugh, L., Tung, R. L., & Van Witteloostuijn, A. (2009). From the editors: Student samples in international business research. Journal of International Business Studies, 40(3), 361–364.

Bernstein, V., Hakel, M. D., & Harlan, A. (1975). The college student as interviewer: A threat to generalizability? Journal of Applied Psychology, 60(2), 266–275.

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Qui˜ nonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. Frontiers In Public Health, 6(1), 149–160.

Bond, M. H., & Hofstede, G. (1989). The cash value of Confucian values. Human Systems Management, 8(3), 195–200.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. Structural Equation Modeling, 9(2), 233–255.

Culpepper, R. A., & Watts, L. (1999). Measuring cultural dimensions at the individual level: An examination of the Dorfman and Howell (1988) scales and Robertson and Hoffman (1999) scale. Academy of Strategic and Organizational Leadership Journal, 3 (1), 22–34.

Dorfman, P., Howell, J. P., Farmer, R. N, & McGoun, E. G. (1988). Dimensions of national culture and effective leadership patterns: Hofstede revisited. Advances in international comparative management (pp. 150–172). London, UK: JAI Press.

Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. Structural Equation Modeling A Multidisciplinary Journal, 6(1), 56–83.

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. Social Psychological and Personality Science, 8(4), 370–378.

Furrer, O., Liu, B. S. C., & Sudharshan, D. (2000). The relationships between culture and service quality perceptions: Basis for cross-cultural market segmentation and resource allocation. Journal of Service Research, 2(4), 355–371.

Gerlach, P., & Eriksson, K. (2021). Measuring cultural dimensions: External validity and internal consistency of Hofstede's VSM 2013 scale. Frontiers in Psychology, 12, 1056–1064.

Hanel, P. H., & Vione, K. C. (2016). Do student samples provide an accurate estimate of the general public? PLOS One, 11(12), Article e0168354.

Hirschfeld, G., & Von Brachel, R. (2014). Improving multiple-group confirmatory factor analysis in R–A tutorial in measurement invariance with continuous and ordinal indicators. Practical Assessment Research and Evaluation, 19(1), 1–13.

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

Hofstede, G. (1986). Cultural differences in teaching and learning. International Journal of Intercultural Relations, 10(3), 301–320.

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

Hofstede, G. (1994). Management scientists are human. Management Science, 40(1), 4–14.

Hofstede, G. (2001). Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations (2nd ed). London: Sage Publications.

Hofstede, G. (2002). The pitfalls of cross-national survey research: A reply to the article by Spector et al. on the psychometric properties of the Hofstede values survey module 1994. Applied Psychology, 51(1), 170–173.

Hofstede, G. (2006). What did GLOBE really measure? Researchers' minds versus respondents' minds. Journal of International Business Studies, 37(6), 882–897.

Hofstede, G., & Bond, M. H. (1988). The Confucian connection: From cultural roots to economic growth. Organization Dynamics, 16, 4–21.

Hofstede, G., Hofstede, G. J., & Minkov, M. (2005). Cultures and organizations: Software of the mind, 2. NEw York: Mcgraw-hill.

Hofstede, G., Hofstede, G. J., & Minkov, M. (2010). Cultures and organizations: Software of the mind. Revised and expanded (3rd ed.). New York: McGraw-Hill.

House, R. J., Hanges, P. J., Javidan, M., Dorfman, P. W., & Gupta, V. (2004). Culture, leadership, and organizations: The globe study of 62 societies. Thousand Oaks, CA: Sage Publications.

Inglehart, R., & Baker, W. E. (2000). Modernization, cultural change, and the persistence of traditional values. American Sociological Review, 65(1), 19–51.

John, O. P., & Srivastava, S. (1999). The big-five trait taxonomy: History, measurement, and theoretical perspectives. L.A. & Pervin & O.P., John. In Handbook of personality: Theory and research, 2 pp. 102–138). University of California Berkeley.

Kirkman, B. L., Lowe, K. B., & Gibson, C. B. (2006). A quarter century of culture's consequences: A review of empirical research incorporating Hofstede's cultural values framework. Journal of International Business Studies, 37(3), 285–320.

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

Kluckhohn, F. R., & Strodtbeck, F. L. (1961). Variations in value orientations. Peterson: Evanston, IL: Row. Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine, 15(2), 155–163.

Kruger, T., & Roodt, G. (2007). Hofstede's VSM-94 revisited: Is it reliable and valid? SA Journal of Industrial Psychology, 29(1), 75–82.

Kuchinke, K. P., & Ardichvili, A. (2001). Work-related values of managers and subordinates in manufacturing companies in Germany, Georgia, Kazakhstan, the Kyrgyz Republic, Russia, and the US. Journal of Transnational Management Development, 7(1), 3–154.

Kyriazos, T. A., & Stalikas, A. (2018). Applied psychometrics: The steps of scale development and standardization process. Psychology, 9(11), 2531–2545 (Savannah, Ga.).

Lippa, R. A. (2001). On deconstructing and reconstructing masculinity–femininity. Journal of Research in Personality, 35(2), 168–207.

Maznevski, M. L., & DiStefano, J. J. (1995). Measuring culture in international management: The cultural perspectives questionnaire (pp. 95–139). The University of Western Ontario Working Paper Series.

McSweeney, B. (2002). Hofstede's model of national cultural differences and their consequences: A triumph of faith - a failure of analysis. Human Relations, 55(1), 89–118.

Minkov, M. (2018). A revision of Hofstede's model of national culture: Old evidence and new data from 56 countries. Cross Cultural & Strategic Management, 25(2), 231–256.

Minkov, M., Bond, M. H., Dutt, P., Schachner, M., Morales, O., Sanchez, C., et al. (2018). A reconsideration of Hofstede's fifth dimension: New flexibility versus monumentalism data from 54 countries. Cross-Cultural Research, 52(3), 309–333.

Minkov, M., & Kaasa, A. (2020). A test of Hofstede's model of culture following his own approach. Cross Cultural & Strategic Management, 28(2), 384–406.

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. Research in Nursing & Health, 29(5), 489–497.

Rokeach, M. (1973). The nature of human values. New York: Free Press.

Schriesheim, C. A., Castro, S. L., Zhou, X. T., & Yammarino, F. J. (2001). The folly of theorizing "A" but testing "B": A selective level-of-analysis review of the field and a detailed leader–member exchange illustration. The Leadership Quarterly, 12(4), 515–551.

Schwartz, S. H. (1994). Beyond individualism/collectivism: New cultural dimensions of values. U. Kim, H. C. Triandis, C. Kagitcibasi, S. C. Choi, & G. Yoon. Individualism and collectivism: Theory, methods and applications (pp. 85–119). London: Sage Publications.

Sharma, P. (2010). Measuring personal cultural orientations: Scale development and validation. Journal of the Academy of Marketing Science, 38(6), 787–806.

Sherer, M., Maddux, J. E., Mercandante, B., Prentice-Dunn, S., Jacobs, B., & Rogers, R. W. (1982). The self-efficacy scale: Construction and validation. Psychological Reports, 51 (2), 663–671.

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

Smith, P. B. (2002). Culture's consequences: Something old and something new. Human Relations, 55(1), 119–135.

Spector, P. E., Cooper, C. L., & Sparks, K. (2001). An international study of the psychometric properties of the Hofstede Values Survey Module 1994: A comparison of individual and country/province level results. Applied Psychology: An International Review, 50(2), 269–281.

Steel, P. (2010). Arousal, avoidant and decisional procrastinators: Do they exist? Personality and Individual Differences, 48(8), 926–934.

Steel, P., & Taras, V. (2010). Culture as a consequence: A multilevel multivariate meta- analysis of the effects of individual and country characteristics on work-related cultural values. Journal of International Management, 16(3), 211–233.

Stevens, C. K. (2011). Questions to consider when selecting student samples. Journal of Supply Chain Management, 47(3), 19–21.

Taras, V. (2019). Catalogue of instruments for measuring culture. Retrieved from htt ps://www.dropbox.com/s/80mk3oebwva1oyk/Culture_Survey_Catalogue.pdf?dl =0.

Taras, V., Kirkman, B. L., & Steel, P. (2010). Examining the impact of culture's consequences: A three-decade, multi-level, meta-analytic review of Hofstede's cultural value dimensions. Journal of Applied Psychology, 95(3), 405–439.

Taras, V., Muchinsky, P., Sarala, R., Avsec, A., Kemmelmeier, M., Aygun, Z. K., et al. (2014). Opposite ends of the same stick? Multi-method test of the dimensionality of individualism and collectivism. Journal of Cross-Cultural Psychology, 45(2), 213–245.

Taras, V., & Orde˜ nana, X. (2015). X-Culture: challenges and best practices of large-scale experiential collaborative projects. V. & Taras & M. A., Gonzalez-Perez. The handbook of experiential learning in international business (pp. 131–149). Houndmills, UK: Palgrave Macmillan.

Taras, V., Rowney, J., & Steel, P. (2009). Half a century of measuring culture: Approaches, challenges, limitations, and suggestions based on the analysis of 112 instruments for quantifying culture. Journal of International Management, 15(4), 357–373.

Taras, V., Rowney, J., & Steel, P. (2013). Work-related acculturation: Change in cultural values following immigration. International Journal of Human Resource Management, 24(1–2), 130–151.

Taras, V., & Steel, P. (2009). Beyond Hofstede: Challenging the ten testaments of cross- cultural research. C. Nakata. Beyond hofstede: Culture frameworks for global marketing and management (pp. 40–61). Chicago, IL: Macmillan/Palgrave.

Taras, V., Steel, P., & Kirkman, B. L. (2012). The times they are a-changin': Improving cultural indices and rankings based on a meta-analysis of Hofstede's dimensions. Journal of World Business, 47(3), 329–341.

Taras, V., Steel, P., & Kirkman, B. L. (2016). Does country equate with culture? Beyond geography in the search for cultural boundaries. Management International Review, 56 (4), 455–487.

Thomson, W. (1883). Electrical units of measurement. Popular Lectures and Addresses, 1 (73). delivered May 3, 1883.

Tsui, A. S., Nifadkar, S. S., & Ou, A. Y. (2007). Cross-national, cross-cultural organizational behavior research: Advances, gaps, and recommendations. Journal of Management, 33(3), 426–478.

Vitell, S. J., Paolillo, J. G. P., & Thomas, J. L. (2003). The perceived role of ethics and social responsibility: A study of marketing professionals. Business Ethics Quarterly, 13 (1), 63–86.

Voich, D. (1995). Comparative empirical analysis of cultural values and perceptions of political economy issues. Westport, CT: Praeger.

Wong, C. S., & Law, K. S. (2002). The effects of leader and follower emotional intelligence on performance and attitude: An exploratory study. The Leadership Quarterly, 13(3), 243–274.

Yoo, B., Donthu, N., & Lenartowicz, T. (2011). Measuring Hofstede's five dimensions of cultural values at the individual level: Development and validation of CVSCALE. Journal of International Consumer Marketing, 23(3–4), 193–210.

Yoo, B., & Shin, G. C. (2017). Invariant effect of individual cultural orientations: An application of CVSCALE. International Marketing Review, 34(6), 735–759.