

# Experimental Methods In Cross-Cultural Management

Ronald Fischer and Johannes Alfons Karl

Experiments have been an integral part of the scientific toolkit. The major advantage of experiments is that they allow us to test plausible cause-effect relations. We humans are intuitive experimentalists; we try to figure out how the world works and how different parts of the world relate to each other. These questions can be of a quite trivial nature and we may learn them early in our life: what happens if I touch a candle; can I eat a stone; what happens if I throw a glass on the floor; what happens if I pull the table cloth at the breakfast table? A simple course of action will give us easy-to-observe answers about causal consequences in real time: I will burn my finger; I cannot eat a stone; the glass breaks; my caregivers will be upset. Other questions are more complex, unfold over longer time periods and require more sophisticated reasoning. What happens if I do not attend lectures on a regular basis; what is the best method for organizing an effective staff meeting; what is the best way to initiate a collaboration with my colleagues in Chile? Here, multiple variables and time scales intersect and it is not possible to establish simple cause-effect relationships.

The core reason for conducting experiments is to test theoretical predictions, especially the causal<sup>1</sup> relationship between variables, or to measure causal processes. There are two major

types of experiments: true experiments and quasi-experiments.

In true experiments, we have control over the variables that we believe are causally involved in creating some change in our dependent variables. We can fully manipulate the variable of interest (give participants no coffee, two cups of coffee or five cups of coffee and then observe the effects on behaviour) and we can randomly assign individuals to conditions (who will drink coffee and who will have no coffee). We have full control over the experimental procedures and assignment of individuals to conditions, therefore the internal validity – the extent to which changes in our dependent variable (e.g. behaviour) are due to changes in our independent variable (e.g. amount of coffee consumed) – is high. However, the extent to which naturalistic conditions can be replicated in such strict ways in an experiment are limited, therefore ecological validity – the extent to which findings are applicable to the real world – are often lower.

Quasi-experiments on the other hand are naturalistic experiments, where individuals are members of a ‘treatment’ condition that is beyond the control of the researcher. Culture is a classic natural or quasi-experiment because people are born into different communities that are likely to influence their behaviour (van de Vijver & Leung,

AQ: Baumert,  
Schmitt, et al.,  
2017;  
Borsboom,  
Mellenbergh, &  
van Heerden,  
2003: Not in  
References

1997). It is quasi-experimental because experimenters cannot randomly allocate individuals to a specific cultural group compared with another and we cannot manipulate or control what specific element of the culture somebody is exposed to (e.g. selective exposure to a high school education system, but not other socio-cultural or socio-economic practices). In this sense, any cross-cultural research by nature is a quasi-experiment (Fischer & Poortinga, 2018).

In summary, experimentation forms part of sense-making processes that humans naturally engage in. The ability to differentiate cause–effect relationships also makes experiments a very valuable method for testing theoretical predictions in a post-positivist and critical realist framework. Yet, cultural processes with their complex temporal and contextual dynamics provide significant challenges for experimentalists. We will provide some discussion and guidance on how to design experiments for cross-cultural research and some of the problems that may emerge and need to be considered. It is a nascent field of enquiry and we hope that our guidelines provide a useful starting point.

### ADDRESSING SOCIETAL NEEDS – WHY WE NEED EXPERIMENTS

Given the complexity of experiments, you may wonder why we should bother with experiments. There are multiple answers that vary in degree, but they all come back to the same answer: because they provide useful information about how the world is likely to work (Falk & Heckman, 2009). We often have competing intuitions about how the world works – should I reward individuals regardless of their input (so as not to upset group harmony) or should I reward the most efficient single employee (to incentivize individual performance)? Is it better to consult employees about forthcoming changes (I risk appearing weak) or should I make decisions because I have all the information (while employees do not have access to all the information) and increase my status as a determined and effective leader? Experiments can help us to sort out which of the two options might be more beneficial in what context, how much of a difference it may make and also when to change course of action. You may now counter that experiments are often highly artificial and do not represent real-world contexts. Survey research or observations might provide much richer context and can tell us when to use which strategy (because we may see that

teams perform better if we use one incentive strategy but not the other). The problem is that both survey and observational studies cannot establish cause–effect relationships (they are providing correlational evidence). Furthermore, in real-world contexts we often face situations where we cannot be sure which of the many interrelated variables is actually the most important for the changes in our variable of interest. Experiments allow us to isolate variables and test whether they can account individually for changes in the dependent variable.

Of course, experiments should be just one method in the larger toolset of a researcher. We need to triangulate our methods and test whether we get similar results when using different methodological tools. Similarly, a single experiment is not going to invalidate results obtained with different methods. But a series of experiments showing a consistent set of results will be informative and can provide important insights for more effective management.

### OPERATIONALIZATION AS THE KEY TO EXPERIMENTS

Experiments are conducted to test some specific theoretical hypotheses. For example, are multicultural teams more or less creative than monocultural teams? In this example, we have theoretical variables and we specify how an independent variable is thought to be causally relevant for changing a dependent variable. Yet, to test this hypothesis we need to translate the theoretical variables into measurable factors, such as responses to questionnaires or employee evaluations. This process is called operationalization and it is key for experiments.

#### *Independent and Dependent Variables*

The independent variable in an experimental design is the variable that the experimenter manipulates (in a true experiment) or that is naturally occurring and thought to causally influence some dependent variable (in a quasi-experiment). The dependent variable is measured by the researcher to see the effect of the independent variable. In experiments the researcher aims to establish causality by manipulating the independent variable and examining the change in the dependent variable. However, the manipulation of the independent

variable (or the natural occurrence of a variable in a quasi-experiment) may also covary with other variables. For example, if we examine the effect of team composition on productivity in a naturally occurring quasi-experiment, the variable of interest (productivity) may for example covary with type of work (more collaborative work structures etc.) or hierarchy levels (lower level teams might already be more diverse). Similarly, our dependent variable may be affected by other variables that were not manipulated (or observed) by the researcher. These extraneous, confounding, nuisance or error variables can affect results and are important to control for. The key variables that are important to consider here are situational variables, person characteristics, experimenter effects and demand characteristics. We will explore them more below, but first we need to discuss the concept of validity.

### ***Internal versus External Validity***

Validity is a core concept in research, indicating the credibility or believability of the results. In measurement terms it is the extent to which a test measures what it is supposed to measure. There are two types of validity in experimental research that are particularly important. Internal validity is the extent to which the changes in the dependent variable are actually due to the manipulation of the independent variable. This is the question of causality. The second type is external validity, indicating the extent to which the results of the experiment can be generalized beyond the context of the experiment. Internal and external validity are often in intrinsic conflict with each other. For example, in a study focused on communication behaviours in intercultural encounters, Pekerti and Thomas (2015) took great care to maximize the internal validity of their study by paying attention to age and gender differences between the samples. On the other hand, their experiment used a standardized setting which is unlike many business contexts, and participants who were unfamiliar with each other were asked to discuss subjective severity of crimes. This is a highly sensitive topic that most might be unwilling to discuss with new acquaintances. Therefore, the setting and topic of discussion might limit the generalizability of findings on communication styles to other contexts. Thus any experimentalist faces a dilemma. On one side, the researcher wants to increase control over the experimental procedure, thereby increasing the internal validity, but on the other side this typically means that the experiment becomes unnatural and many

real-world relevant processes and variables are being excluded from the experimental context, which then lowers the external validity (extent to which the experiment still describes real-world processes). Therefore, the balance of internal vs external validity is one of the key decisions an experimental researcher must make.

### ***Testing Internal Validity – The Importance of Manipulation Checks***

One option to test the effectiveness of an experimental design is to check whether the manipulation of the independent variable was effective. These so-called manipulation checks are important to determine the quality of the experiment. The inclusion of manipulation checks is particularly important in the context of quasi-experimental studies, in which the experimenters have no control over the strength of the independent variable. Therefore, they need to ascertain that the independent variable is operating in the way that it is theoretically expected. In cross-cultural quasi-experiments, manipulation checks are often done using validated measures of cultural values (e.g. Hofstede, 1980), norms or self-construals (e.g., Singelis, 1994), but individual-level measures can be used to check whether samples are indeed as culturally distinct as theoretically expected. For example, a recent study examined whether culturally mixed groups would outperform culturally homogeneous groups in computer-mediated group decision tasks and used individualism–collectivism measures to test the cultural orientation of participants (Li, Rau, & Salvendy, 2014). One of the pitfalls of this particular approach is that (a) it strongly depends on the definition of culture that a researcher adopts and (b) it raises the question whether the measures that are used for the manipulation accurately check operationalization of the relevant cultural dynamics (Fischer, 2012).

A second and often underappreciated problem is that many cultural dimensions covary. For example, individualism–collectivism often covaries highly with power distance and national income levels (see Hofstede, 1980). If a researcher were to measure only individualism–collectivism, they would likely find significant differences between two or more samples. However, the differences in the dependent variable might be driven by the unmeasured third variables of power distance or national wealth that covary with individualism–collectivism (Smith, Fischer, Vignoles, & Bond, 2013). Overall, properly conducted manipulation checks are an important tool to test whether the experiment worked as intended.

### **Randomization**

The key element of a true experiment is that individuals are randomly assigned to treatment conditions (variations of the manipulated independent variable). In cross-cultural experiments, this is often not possible. For example, we cannot randomly assign individuals to a culture. In such quasi-experiments we lack control of the assignment of individuals to conditions and, therefore, any claims about causality are weakened and alternative explanations are more difficult to control. If we conduct a within-culture experiment (e.g. compare the effects of team composition on performance), it might be possible to assign individuals randomly to treatment conditions. But even here, it is often difficult to assign people randomly because of naturally occurring variations that are not random (age, gender, professional specialization, etc.). One of the most important tasks for experimenters is to identify these biases or error variables and control for their effects in the experiment.

### **EXPERIMENTAL BIASES**

Biases introduce error and challenge the validity of the findings in our experiment. In terms of internal validity, we can distinguish two major forms of error: random error, which is unsystematic; and systematic error that is due to some meaningful third variable. For example, if we conduct a study on the effect of group composition on work performance, some people might be more or less productive on that particular day (random variation, because it is not a systematic difference) and some people might be working daily in very diverse teams (systematic error, because it is a stable variable between individuals that can be controlled).

The interesting question is what effects random vs systematic error might have on the results. As a general guideline, we can expect that random error will increase the general noise in our study and therefore limit our ability to adequately detect an effect (see the upper part of Figure 7.1). However, the mean effect size (the relative difference between conditions) in an experiment remains unchanged if the error is completely random. Increasing our sample size is an effective strategy to increase power to detect effects if there is too much random error in an experimental setup. One important question for researchers is to determine an adequate sample size, for example through a priori power analysis. In contrast to random error,

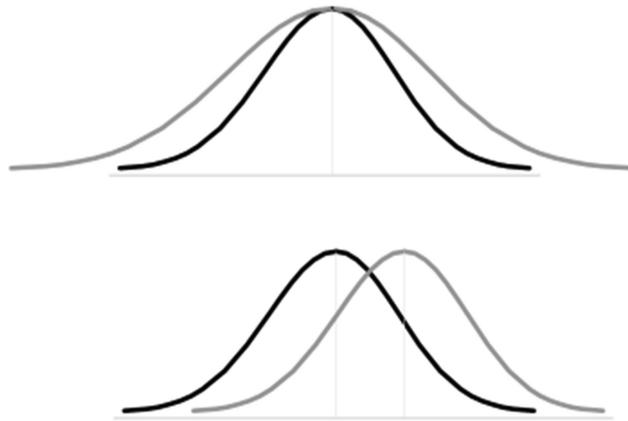
systematic error can shift the mean or change the substantive conclusions of our study (see the lower part of Figure 7.1). Here, a third variable may influence results and shift the means in one or more conditions.

### **Situational Differences and Standardization**

The first set of variables that may systematically influence results is situational differences. For example, if the experimenter wears a formal dress on the first day but a casual outfit on the second day, or may give quite firm and serious instructions in one setting but appears quite relaxed and chatty in a second run of the experiment, this will induce both random and systematic error into the experiment. Similarly, the environmental conditions need to be carefully controlled. Team performance will be affected if there is noise in one setting compared with the other, or if the ambient temperatures are much higher in one context compared with another. The experimenter has to make sure that the experimental conditions are as similar as possible, using carefully worded scripts for instructions and minimizing any other variation in the experimental context. This is often called standardization (Fischer & Poortinga, 2018). It can be challenging when conducting experiments across different cultural contexts, because temperature and noise levels as well as many other ambient features of the experimental context vary systematically across the world. The key concern here is to standardize the procedures, context and conditions as much as possible to rule out both random and systematic error.

### **Person Characteristics and Confounds**

A second important issue to consider is person characteristics. Are there systematic differences between individuals that may invalidate or challenge conclusions that an experimenter draws from an experiment? In cross-cultural research, income and education are often potent challenges to any claims about cultural differences (Hofstede, 1980; Smith et al., 2013). These are often called confounding variables because they may either covary with the experimental manipulation or independently influence the dependent variable and thereby confound the conclusions that might be drawn from the experimental outcomes. This is particularly relevant for quasi-experiments because participants are not randomly assigned to



**Figure 7.1** The effects of random and systematic error on distributions

*Notes:* The upper distribution shows the effect of random error (upper curve) on a single population – random error increases the variance but does not affect the mean; the lower distribution shows the effect of systematic error (the right-most curve) on a single population – systematic error typically changes the mean but not the variance

conditions and many other variables are likely to covary with the cultural or ethnic background of participants.

Another crucial issue to consider in relation to participants is selection bias: do people self-select to participate in the experiment? Individuals who volunteer for participation may be different in important ways from the general population to which an experimenter might want to generalize the findings to (Falk & Heckman, 2009). Refusal to participate is the opposing tendency and may also create problems. Some topics might be particularly sensitive in one cultural context and lead to increased refusal, which leads to systematic differences.

### ***How to Control Person and Confounding Variables: Blocking***

In experiments, we want to manipulate the effect of one variable on key outcomes. But often there are other variables that influence our outcomes that we want to control. These so-called nuisance variables can be controlled either through randomization or through blocking. The idea of blocking is that instead of assigning individuals randomly to conditions, we purposefully distribute individuals to experimental conditions depending on the variables that we want to control for. Gender is one of the most commonly used blocking variables – we typically want to make sure that we have equal numbers of males and females in

all experimental conditions. Blocking works easily with a single independent variable and a single blocking variable. In organizational settings, we often encounter multiple confounding variables that we want to control for. In these situations, various extensions of the simple block design can be used, such as Latin Squares. Block designs are an elegant mechanism to control variation across experimental conditions that are not of interest to the experimenter (controlling error variance), which helps to increase internal validity and reduces the plausibility of alternative explanations of the results.

### ***Demand Effects***

One of the key advantages of an experimental design is that individuals assigned to conditions typically do not know what the other conditions are. Yet, there are still challenges. Demand effects, for example, refer to clues that participants pick up from experimental procedures, the experimenter or their surrounding that convey the true purpose to the participants. This then may induce motivations in the participants to conform to these inferred demands. For example, researchers might be interested in the extent to which people may fake their personality scores in job interviews and whether such tendencies differ across cultures (Fell & König, 2016). A classic manipulation is to ask participants either to pretend to apply for a job or to answer the questionnaire honestly. Yet,

the instruction to answer honestly may already induce some expectations that then influence the results.

The Hawthorne effect is another classic effect in the organizational literature that is important to consider – just by being aware that you are being included in a study is likely to change work behaviour (McCambridge, Witton, & Elbourne, 2014). Therefore, it is important to include adequate control conditions. To the greatest ethical extent possible, researchers should minimize the ability of participants to guess the real purpose of the study and the relevant design features. This is often called a single blind trial – efforts are made to obscure as much as possible the true purpose of the experiment.

### ***Experimenter Effects and Double Blind Experiments***

In single blind trials, even if participants are unaware of the study design features, the experimenter often knows which participants are assigned to what condition and the experimental predictions (e.g. including specific hypotheses being tested). This raises a separate problem: the experimenter may unconsciously differ in their treatment of the participants and thereby influence the results in subtle ways. This was beautifully demonstrated in an experiment by Doyen, Klein, Pichon, and Cleeremans (2012), in which they tried to replicate a classic priming study by Bargh, Chen, and Burrows (1996). The original study had shown that if participants see words associated with old age, they will work more slowly (e.g. implying that they behaved more like an elderly person). In Doyen et al.'s (2012) study, the participants were assigned to either an age or a control priming condition. Importantly, experimenters were also randomly assigned to conditions, believing that the priming would either increase or decrease walking speed. Indeed, the researchers did not find a participant prime effect, but an effect of experimenter manipulation. Subjects walked slower if the experimenter believed that priming would decrease walking speed. This suggests that the experimenter subtly influenced the behaviour of participants. Therefore, so-called double blind trials in which participants and the experimenter are unaware of the conditions that each participant is assigned to are an important feature of strong experimental protocols.

In cross-cultural research this might be particularly important because of implicit biases and stereotypes about cultural and ethnic differences that researchers may have and influence how people will perform (Cuddy et al., 2009). For example,

high-status ethnic groups are universally seen as more competent but typically less warm, whereas low-status ethnic groups are often seen as less competent but warmer (Cuddy et al., 2009). Such stereotypes that are widespread may now influence how people perform in experimental settings (Dar-Nimrod & Heine, 2011), which makes it necessary effectively to control for these experimenter effects. At the same time, it should be obvious that blinding researchers to the quasi-experimental condition of ethnicity or cultural background is pretty difficult to achieve. In these cases, one option might be to use experimenters who are unaware of the purposes of the experiment to control at least the direction of possible stereotype influences.

### ***Within- and Between-Subject Designs***

One important tool to increase power is to decide on the design of the study. Between-subject designs allocate individuals to different conditions, therefore every participant is only included in one experimental condition. This has the advantage that each participant is only exposed to one condition of the manipulation, which reduces guessing and demand characteristics. For example, we could test whether individuals working in a monocultural team compared with a multicultural team have higher productivity. However, we have only one score per individual and many individual differences may influence the dependent variables. Hence, random error due to individual differences increases and power typically decreases (see Figure 7.1).

Alternatively, we could test the same individuals across all levels of the independent variable. This is called a within-subject design, because the same individuals will respond to all levels of the independent variable. This is often a more powerful approach to control individual differences. For examples, we could test the same individuals in different team compositions and repeatedly record their work performance. This allows us to account for individual differences, decreases the noise and increases the power of our study. However, this also raises a number of problems such as order effects: the earlier administration of a test condition may influence the performance on a later condition.

One option is to counterbalance the administration of conditions to individuals. One group of individuals may work in a monocultural team first, then in a multicultural team, whereas the pattern is reversed for the other group. However,

within-subject designs may not work if there are strong carry-over effects from one condition to the other, if the exposure to one experimental condition does expose the nature and purpose of the experiment, or if the treatment cannot be counterbalanced (you cannot counterbalance stable person characteristics such as age or gender).

In summary, experiments can be run as both between- and within-subject designs. In between-subject designs the participants experience only one condition of the experimental manipulation, whereas in within-subject designs, the same participants are exposed to all conditions of the independent variable. It is possible to include both within- and between-subject variables in so-called mixed effects designs. In cross-cultural research, cultural background typically can only be used as a between-subject variable, but other experimental procedures can be manipulated within individuals and therefore can be treated as within-subject variables.

## VALIDITY, EQUIVALENCE AND BIAS

Considerations about experimental design, blinding and control are relevant to monocultural and cross-cultural studies. Nevertheless, cross-cultural studies raise unique issues, such as equivalence and bias. Equivalence and bias concern the questions of whether we can trust the measurement scores or whether there was some cultural bias in the procedures that challenges the interpretation of results. These issues have been most explicitly discussed in the context of survey methods, but the same principles are relevant for experiments (Fischer & Karl, 2019; Fischer & Poortinga, 2018; Fontaine, 2005). Whenever we want to draw some conclusions about cultural differences or similarities, we have to answer some simple questions about the constructs and measurement process. These questions need to be asked for both the independent and dependent variable, but the specific issues may vary somewhat by the status of our variable (independent vs dependent). This is due to the fact that the experimenter typically manipulates the independent variable experimentally (which requires slightly different considerations), but measures the dependent variable.

The first question is whether the intended construct does exist in each of the cultural groups and whether it can explain behavioural differences across cultural groups and conditions. For both the independent and dependent variable,

we need to discuss whether a concept ‘exists’ and is relevant for participants. This is the issue of functional equivalence (Fontaine, 2005). For example, if we want to manipulate the effect of autonomy on job performance, we need to consider whether the concept of ‘autonomy’ exists and how participants in each group might conceptualize autonomy. For the dependent variable, we face similar concerns: is a particular construct relevant for capturing the behavioural differences that we aim to produce through the manipulation of the independent variable? Therefore, we need to consider whether our dependent variable is relevant for understanding behavioural changes in each cultural context.<sup>2</sup> For example, if we were interested in the effects of cultural diversity on creativity, we might need to ask what counts as cultural diversity in each and every cultural setting (is religious diversity part of our understanding of cultural diversity?) and what is considered as creativity (how does each cultural group understand creativity?).

A second question that is tightly linked to functional equivalence is structural equivalence. This question is concerned with the operationalization of the variables of interest: can we use the same manipulations for the independent variable and can we use the same items or behavioural indicators to measure our dependent variable? For example, if we want to study the effect of anxiety on test performance we need to consider what situations may induce anxiety in different cultural contexts. In many Western settings, it is often embarrassing and nerve-wrecking to sing in front of others, whereas in many other contexts, the practice of singing in front of others is a favourite pastime (Karaoke) and therefore much less anxiety inducing. Similarly, if we wanted to measure anxious responses, they need to be captured in culturally appropriate ways. ‘I feel blue’ may indicate some form of loneliness or depression in English, but in German it implies intoxication, and in Chinese it may refer to immortality. Therefore, the operationalization of constructs needs to be sensitive to the specific conditions and meanings in each cultural group.

A third question concerns equivalence of both the specific changes induced by the manipulations and the sensitivity of measurements of individual indicators of the dependent variable. For the independent variable the question is whether an experimental condition induces the same extent of behavioural change in the dependent variable. For example, does team composition result in the same behavioural activation in a sample of South African workers (who live in a very ethnically diverse society) compared with a sample of South Korean workers (who live in a very homogeneous

society and might therefore have less experience of a culturally mixed team).

Similarly, does the measurement of our dependent variable translate across cultural groups? Can we compare the responses on a Likert scale directly across groups? This is an issue of the underlying metric: does a response of 3 on a rating scale that ranges from 1 to 5 indicate the same behavioural expression of a behavioural trait across different cultural groups? From a measurement perspective, we have to consider whether changes in the intended construct are associated with equal changes in the observed constructs in each group and whether there are baseline differences between the different cultural groups. These questions are investigated through the analysis of different statistical parameters, and more technical discussions can be found elsewhere (Fischer & Karl, 2019; Fontaine, 2005).<sup>3</sup> We focus on the general principles to illustrate their relevance to experiments.

To give an example, the first question is whether our experimental treatment induces the same amount of change in our dependent variable. Is being part of a multicultural team inducing the same amount of behavioural change in each group? Similarly, are our dependent variables sensitive enough to pick up the changes in the independent variable and lead to the same amount of observable (measurable) changes in our dependent variable? For example, if we asked people verbally how productive they were in their team, they might have different expectations due to their previous work experience. Hence, a self-report scale might not indicate the same amount of change due to differential expectations, leading to different sensitivity in each cultural measurement context. If this condition is met, researchers call it metric equivalence.

However, there might still be baseline differences. For example, workers might have different expectations about productivity standards, which influences their ratings independent of the manipulation. This may result in the two groups having different baselines of performance ratings. If no such baseline differences exist, then researchers argue that conditions of full score or scalar equivalence have been met.

### ***Equivalence – Implications for Cross-Cultural Experiments***

The issue of equivalence is of fundamental importance for cross-cultural studies. Unfortunately, it is often ignored in cross-cultural survey research (Boer, Hanke, & He, 2018) and

this situation is worse in the nascent field that uses cross-cultural experiments (Fischer & Poortinga, 2018). Lack of functional equivalence signals that there is construct bias, implying that the constructs do not have the same meaning and function in all the cultural samples studied. If there is no functional equivalence, we cannot compare any findings across cultural groups. For example, a study of the use of social media on work effectiveness is not going to be comparable if social media is blocked or largely unavailable to participants in one cultural context compared with another.

Construct equivalence raises interesting questions from a comparative perspective. If we use different operationalizations of theoretical variables, we might be able to study the theoretical processes in each cultural context individually, but it might be problematic to compare scores directly across cultural contexts. Is the induction of social anxiety through a public singing task (e.g. in Western settings) compared with a public speaking task (e.g. in non-Western settings) leading to comparable changes in the dependent variable which mean that we can directly compare the results? Similarly, if we were to use different indicators for measuring our dependent variable, could we directly compare the results across cultural groups?

If functional and structural equivalence are met, we have more opportunities to compare results directly. If metric equivalence is met, we can compare score patterns – for example, are the treatments of the independent variable leading to more or less change in the dependent variable? However, the direct comparison of scores is still problematic (because there might be baseline differences). Fontaine (2005) suggested that it is possible to compare results under conditions of metric equivalence as long it is made clear that the results apply only to the specific operationalization of the crucial variables. For example, instead of making generalizations about the role of social anxiety on creativity, we could say that public speaking (compared with a reading control condition) leads to a lower number of responses on the alternative uses task (as a measure of creativity) in both culture A and B. Note here that the interpretation stays at the level of the manipulated (speaking vs reading) and measured (number of responses) variables instead of making claims about anxiety and creativity. Only full score or scalar equivalence allows us to extrapolate to the level of the theoretical construct. This is an important interpretational difference, which unfortunately is often missed in empirical research (see Boer et al., 2018; Fischer & Poortinga, 2018).

## PROGRESS AND CHALLENGES IN CROSS-CULTURAL EXPERIMENTS

### *Manipulating the Independent Variable: Priming*

One type of experimental design that has enjoyed popularity in psychology is priming. The main idea of priming is that the researcher identifies a causally relevant cultural context variable and then experimentally manipulates the salience of this variable. Participants are randomly allocated to conditions and are asked to engage in a brief, ostensibly unrelated, activity during which particular psychological concepts, knowledge or motivational goals are activated. Primes commonly involve some form of individualism–collectivism or independent–interdependent self-construals (Oyserman & Lee, 2008), aimed at increasing the salience of individual or group orientation in people's minds. Following the prime, participants will complete the core task of interest. The idea is that the activated concepts from the prime are carried over to the next task and influence people completing the second task. Overall, priming uses a classic between-subjects design, under the assumption that priming targets the active ingredient of culture, to explain previously reported differences between groups in quasi-experimental studies.

For example, Fu, Zhang, Li, and Leung (2016) used priming with Chinese managers to investigate the effect of culture mixing on acceptance of organizational change. They found that respondents who were primed with culture mixing were more willing to accept a hypothetical drop in their salary than those primed with Chinese culture, if they also had a low need for cognitive closure. The researchers concluded that enhanced change acceptance might be motivated by open-mindedness. This unpackages 'culture' by identifying what cultural aspect is relevant for a specific behaviour and then experimentally examines whether it can account for behavioural changes across cultures.

One of the critical questions for priming is external validity, specifically whether priming can simulate the complexities of culture. Some proponents of priming have even claimed that priming can be used to turn Chinese into Americans and vice versa (Oyserman, Sorensen, Reber, & Chen, 2009). If this were true, then most intercultural trainers would be out of work because intercultural training should be easy. Appropriate responses in novel cultural settings should be easily triggered through using the appropriate situational cues. Yet, the large body of acculturation and cross-cultural training research

clearly demonstrates that this is not the case. The heart of the issue is that culturally relevant cues are not easily decoded and meaning needs to be learned through a complex process of acculturation and socialization. This is probably most eloquently summarized by Fiske (2002):

Construct accessibility is a transient effect that cannot be equated with the enduring, objective social entity that is culture. Some individuals are flexibly capable of participating in multiple cultures, of course. But does a person's culture change in response to questions such as 'How are you different from other people?' Does reading that sentence change a Chinese communal farmer into a cowboy? Sort of, a little bit, for a moment? No. If it did, then IND and COL scales would alter culture instead of measuring it. Priming does not change institutions, practices, or systems of communication and coordination. Priming does not affect socially constituted entities, relations, and practices in relation to which a person lives: rodeos, poker, cattle brands, Colt 45s, and gun-fights. If one does not know Wyatt Earp and the OK Corral, they cannot be primed. Mere accessibility can hardly be an important factor mediating the effects of these constituents of culture on the psyche, unless one postulates that all humans have cognitive representations of all significant aspects of all cultures. (pp. 80–81)

Therefore, if we assume that primes are understandable to more than one cultural group, but yet equally effective for producing behavioural change, then most likely we are not dealing with a cultural process, but rather a social effect that is common to all human groups studied. Such experiments can be very informative for management scholars by identifying how situational cues can change behaviours (e.g. turning individuals to act more prosocially or engage in more creative behaviour), but it probably does not tell us much about the cross-cultural differences of interest.

A different approach could be to understand cultural dynamics within cultural groups. For example, the widespread use of corruption across the world is of great concern and importance for management scholars. For example, countries with high levels of corruption can be easily identified, yet the individual-level conditions that motivate individuals to engage in these behaviours are not well understood. It is possible to use priming to study culture-specific mechanisms. A recent study (Fischer, Ferreira, Milfont, & Pilati, 2014) examined the extent to which Brazilians are more or less likely to report intentions to engage in corruption behaviour depending on whether

they were primed with images that (negatively) portray corruption or focus on cultural icons that creatively and unconventionally solve problems through bending social norms (so-called *malandros*, individuals who live on the margins of formal society, but are associated with many positive qualities such as samba and soccer). Corruption primes, even though they were negative in nature, increased the willingness to engage in corrupt behaviours in a general population sample, due to decreasing the thresholds (e.g. corruption is negative, but it is commonly practised and hardly punished, implying it is a social norm). Being primed with cultural icons also increased intentions to use these behaviours, but only among those individuals who identified greatly with Brazil (most likely due to the activation of associated positive concepts such as samba and soccer). Hence, the activation of culturally salient and meaningful primes can explain when and why people might engage in corruption behaviour. These patterns would not be meaningful and understandable to outsiders who have no familiarity with the specific cultural context cues.

A second theoretical question concerns the effectiveness of priming. Many priming stimuli do not work as well in Asian contexts as in Western contexts (e.g. Oyserman & Lee, 2008). This raises interesting questions about the experimental designs. It might be possible that the primes did not show structural or metric equivalence, which is an operational issue. Alternatively, the primes may trigger different responses in Asian samples compared with Western samples, implying different theoretical processes. These might be operating at the level of constructs (e.g. the constructs are not functionally equivalent) or they may indicate true cultural differences of a substantive nature. These questions have not received sufficient attention to allow us to draw any conclusions in either direction. In summary, we believe that priming studies have potential in illustrating culture-specific processes for cross-cultural management scholars, yet the current literature has focused mostly on priming studies that do not capture the rich cultural context implied by definitions of culture. Hence, the operationalization of the key variables in these priming studies has been deficient.

### ***Assigning Individuals to Conditions: Team Diversity Studies***

Team diversity studies focus on recreating culturally mixed or intercultural teams by recruiting participants with specific cultural backgrounds and

assigning them in a manner that leads to the desired levels of cultural mixing. For example, Li, Rau, and Salvendy (2014) created culturally mixed groups by gradually altering the composition of the groups from five Chinese locals in one group to four American and one Chinese local per group. The researchers were interested in the impact of cultural mixing on decision quality and speed and found that cultural mixing increased quality, but also time needed to find a consensus. Team diversity studies allow for high internal validity because all team members are unfamiliar with each other and the degree of cultural mixing can be accurately controlled. Nevertheless, these studies might not have high external validity, as the mixture is artificially induced and may not reflect team diversity in real organizations.

### ***Using Naturally Occurring Variation: Field Experiments***

Even within cultures, individuals often operate across different situational contexts that may influence their behaviours, such as altruism and social cooperation. Going out of the way to help others or to cooperate when this cannot be enforced or rewarded is a key organizational citizenship behaviour that is of importance for organizational survival and effectiveness (Farh, Zhong, & Organ, 2004). Field experiments can provide powerful new insights into how collective behaviour and situational context (presence or absence of behavioural cues) can shift behaviour (Krátký, McGraw, Xygalatas, Mitkidis, & Reddish, 2016).

For example, anthropologists have long noted that specific situational cues might increase or decrease altruism in naturalistic settings. These variations are often associated with organized institutional structures (such as religion), but changes in altruism may also occur in other social settings where individuals behave in ritualistic ways (Islam & Zyphur, 2009). One of the best studied examples is the positive effect of synchronized behaviour (e.g. singing, marching, dancing) on feelings of connectedness and anonymous monetary contributions to the group (Fischer, Callander, Reddish, & Bulbulia, 2013; Mogan, Fischer, & Bulbulia, 2017). Some of these rites and ritualistic features have been studied in the context of organizational culture research, mainly through ethnographic observation and more recently through surveys (Fischer et al., 2014). Given the importance of symbolic action and rites more generally in the functioning of

organizations across cultures, this is an area ripe for investigation.

## WHAT IS THE FUTURE OF EXPERIMENTS IN CROSS-CULTURAL RESEARCH?

### *Replication*

While we often assume experiments to be the gold standard of scientific enquiry into causality, this can lead to overconfidence in reported experimental results. In 2015 a paper was published by the Open Science Collaboration that reported on an attempted replication of 100 studies from top psychology journals (Open Science Collaboration, 2015). The authors found that they were only able to replicate a third of the original effects using stringent criteria or half of the original reported effects using more lenient criteria to assess successful replication. This article has clearly struck a nerve in the psychological community and has been cited more than 3,000 times (as of May 2019). The large-scale failure to replicate findings in the psychological literature has been termed the replication crisis. The low replicability is not unique to psychological science and might be even worse in other fields, such as oncology (Begley & Ellis, 2012). Low replicability can stem from a number of sources, such as low sample size, small effect sizes, data dredging (also known as P-hacking; for an interesting example see: <http://shinyapps.org/apps/p-hacker/>), conflicts of interest between individuals and institutions, large numbers of scientists working competitively without combining their efforts as well as the file drawer effect. The file drawer problem (non-significant findings are not published, but instead filed away) is especially problematic as it skews the available literature in an area. Researchers might identify a gap in the literature without the knowledge that this area has been researched previously, but the research was not published due to non-significant results (Rosenthal, 1979). Different authors have proposed solutions to this problem, ranging from repeated attempts at replication to determine the true effect (Maxwell, Lau, & Howard, 2015) to making replication an integral part of PhD theses (Everett & Earp, 2015). In the management sciences, the problem of low replication rate of studies might limit the cumulative advancement of the field (Hubbard, Vetter, & Little, 1998) and the implications of this possibility are gaining attention (Tsang & Kwan, 1999). We believe that replication is an essential element

of the scientific process, especially when done in a rigorous fashion (for a template of replication see: Brandt et al., 2014). Nevertheless, replication represents only one building block necessary to increase the reliability of scientific findings.

### *Pre-registration versus Exploratory Research*

Successful replication relies on replicable studies. While currently journal articles report which methods were applied, they are often sparse on exact details, which increases the hurdle for successful replication. To increase transparency, accountability and replicability in science, pre-registration of studies has risen in popularity since the replication crisis and some journals now require pre-registration for publication.

Pre-registration, in its simplest form, may simply comprise the registration of the basic study design before it is conducted. More commonly, pre-registration also includes a detailed pre-specification of the study procedures, outcomes and statistical analysis plan. Take for example our team diversity experiment. If we wanted to pre-register this experiment, we would need to specify our hypotheses, study design, sampling procedure and planned analysis before we start. Pre-registration is beneficial not only for the scientific community, for example by combating the file drawer problem, but also for the individual researchers as questions about hypotheses and analysis are solved before data are collected. Support for study pre-registration is increasing; websites such as the Open Science Framework (<http://osf.io>) offer services to pre-register studies. While replication of existing studies has gained attention in management sciences, pre-registration is less common.

It is important to emphasize that this focus on pre-registration should not come at the cost of exploratory research. The real issue is that exploratory research is often presented in publications as if it was confirmatory. Probably many of us have faced this dilemma when a particular finding emerged in our data which in hindsight can be expected and explained by a different theoretical paradigm. Management journals typically prefer hypotheses instead of research questions, and it is important to realize that a thorough exploration of data without theoretical blinders and paying attention to what data can tell us about a problem is probably as important as hypothesis testing. Many of the most important discoveries in human history were caused by accidents (e.g. penicillin, Post-it notes). In other words, we urge researchers

to be clear about what part of their study is truly theory testing and pre-register those elements, but also to explore the data to find unexpected or surprising findings that can provide us with more insights into how organizations are operating.

### ***Essentializing vs Cultural Dynamics***

Much of cross-cultural research assumes a static perspective on culture – individuals socialized into a specific socio-cultural–economic environment will have a certain mindset, values and beliefs which then translate into specific behaviours that collectively can be studied by researchers and inferences can be drawn about the population-level differences. This ignores the increasing complexity of modern nation-states, multiple sources of cultural influence within and across cultures (for an interesting example see: Ferguson et al., 2016), remote acculturation and increasing diversity within organizations because of migration, intermarriage and shifting population dynamics. The point here is that cultures are dynamic and ever changing. This provides a potent challenge for any cross-cultural study, but it is of particular importance for experimental designs because the aim is to examine causal relationships, which may not be temporally stable.

### ***A Cross-Cultural Experimentalist's Checklist***

We wrote this chapter to provide some introduction to experimental designs and showcase their applicability for cross-cultural management researchers. In order to help researchers develop useful experimental studies, we finish off by providing a generic checklist that covers some of the key points. Specific projects may need further elaboration or attention to some of these points.

#### ***State your question and theoretical background clearly***

Experiments are most useful if a researcher has a clear expectation of causality of their theoretical model. From this background a clear research questions should be defined that delimits the scope of the research.

#### ***Formulate your hypothesis and identify key theoretical variables***

Theoretical statements need to be translated into testable hypotheses. Which independent variables

are supposed to cause what changes in a dependent variable and through what mechanism? The best hypotheses clearly specify the independent and dependent variables, the direction and magnitude of change, and any boundary conditions. What is the role of culture in your hypothesis? Carefully consider whether you expect the hypotheses to be identical across cultures or not.

#### ***Identify your cultural groups based on cultural variables of interest***

Comparison cultures should be selected based on the theoretical foundation of the research rather than convenience samples. How does the theoretical process work within and between cultural groups? In cross-cultural research, it is important to consider explicitly what cultural variables are of importance as either conditional moderator variables or as mediators.

#### ***Identify the feasibility of an experiment***

Once the theoretical question and the cultural dimensions have been identified, it is important to consider whether it is actually feasible to conduct an experiment. For example, many questions cannot be easily addressed with a traditional experimental paradigm because of the unethical nature of manipulating ‘culture’ (e.g. raising randomly selected individuals in different cultural contexts). Yet, it may be possible to use other experimental techniques (e.g. priming, quasi-experimental or field studies) to investigate causal processes as predicted by your theory.

#### ***Operationalize your variables***

Theoretical variables need to be translated into measurable variables in an experimental context. How can you operationalize your theoretical variables in the specific experimental context? Make sure that your theoretical variable and its experimental manipulations or measurement are logically consistent with each other. Include a control group that allows you clearly to identify experimental effects. Consider whether your manipulation of the independent variable and the measurement of dependent variables are possibly picking up unwanted or unintended other theoretical processes. A very important issue to consider is whether the operationalization may introduce cultural biases into your experiment. Discuss the operationalization of your variables with cultural experts.

### *Design the experiment and write your experimental protocol*

Once you have specified your hypotheses and operationalized your key variables, you need to write an experimental protocol that clearly specifies all the key steps of your experimental design. Is it a within- or between-subject design? How do you deal with confounds? Do you use blocking or counterbalancing? What are experimenters doing and when? What materials are needed? How are individuals assigned to experimental conditions? What is the cover story? Provide criteria for participant exclusion (how do you decide whether a participant needs to be excluded or not?). Conduct a power analysis to decide how many participants are needed in each condition to test your hypotheses adequately. Provide sufficient detail so that other people can take your design and run the study independently of you.

### *Obtain ethical approval from the relevant authorities*

Once you have your experimental protocol ready, you need to gain ethical approval from the relevant authorities before you can proceed with your study. In the ethical approval review process, some concerns or issues with the design or the sampling might be raised. Cultural differences in conducting research may need particular attention (e.g. signed consent might be inappropriate in some cultural contexts). Talk to the head of the appropriate ethics committee or review board to understand and clearly address all relevant issues. An additional issue might be that not all countries or locations where you are aiming to collect data may have functioning ethical review boards. Consult your ethics board about what to do in those cases.

### *Pre-register your study*

We strongly recommend pre-registering your study protocol and hypotheses. The advantage of experiments is a careful testing of causal relationships. Any conclusions are strengthened if you can demonstrate that you followed your own protocol. Of course, it is legitimate to conduct exploratory analyses, but these should be clearly indicated as such. An added challenge in cross-cultural research is that all measures need to be tested for equivalence. Equivalence is always sample dependent, therefore it is necessary to test measurement invariance in every sample and at each stage. For your pre-registration, it would be important to identify how you decide on measurement invariance and what criteria you will use for identifying and eliminating bias.

### *Conduct the experiment and document all deviations*

Run your experiment. Follow your study protocol. If necessary, run a pretest to test whether the protocol and design are working as intended. When running your study, document any deviation from the study protocol and any other information that may be useful in interpreting the study results (e.g. unusual participant characteristics, noise or accidents). Especially in cross-cultural settings and field studies, unexpected events may happen that can qualify the quality of the study. Carefully note anything that might be relevant for later consideration.

### *Analyse and write up your results*

In your study protocol and pre-registration, you should have specified how you are going to analyse your data. Run your manipulation checks to see whether your experimental procedure was effective. An added complexity in cross-cultural research is the issue of equivalence and bias. It is important that you test whether your measures worked equally well in each cultural setting (see the pre-registration part). Once you are satisfied that your measures are working adequately in all the samples, you can test your theoretical hypotheses. In case your measures fail the equivalence test, you have to address the theoretical implications. Failures to find measurement invariance are highly informative about cultural dynamics. In our opinion, these failures are often more exciting and fascinating in what they reveal about cultural processes (possibly pointing towards cultural relativism) than a straightforward comparison of means.

In summary, we hope we have provided an informative and instructive overview of experimental methods and their value for cross-cultural management. As we have outlined above, experimental methods have desirable properties that should make them appealing to cross-cultural researchers. It is important to have experiments as one important tool in a larger toolkit to complement other methodological tools.

## **ACKNOWLEDGEMENTS**

We are grateful for feedback from the editors and students enrolled in the MSc in Cross-Cultural Psychology at Victoria University of Wellington, in particular Samuel Twitchin.

## **Notes**

- 1 Causality can be defined in a number of ways. Here we use some working definitions common

across psychology. Specifically, we assume that X causes Y if (a) X and Y covary, (b) X precedes Y and (c) Y is absent if X is absent (Baumert, Schmitt, et al., 2017; Borsboom, Mellenbergh, & van Heerden, 2003). It is important to note that there might be more than one cause for a given variable, which may not occur in the same instance. For example, a fire may break out if (a) a candle is lit, (b) left unattended, (c) surrounded by flammable material and (d) something topples the candle. In this case, a number of variables need to come together for a particular reaction to occur.

- 2 It is important to note that our treatment of functional equivalence for the dependent variable goes beyond the standard discussion of functional equivalence in the psychometric literature. Our treatment of functional equivalence is a conditional question that depends on the manipulation of the independent variable *and* the theoretical nature of the dependent variable itself. In standard psychometric discussions, the theoretical status of the variable to be measured is the core concern. However, for experimentalists an additional concern is the relationship between the independent and dependent variable because the key question is the change in the dependent variable depending on the manipulated status of the independent variable. Therefore, the dependent variable in each cultural group has to be functionally equivalent and sensitive to the manipulations of the independent variable. To provide an example, if we were interested in the effect of repetitive behaviour (Karl & Fischer, 2018) on the reduction of anxiety in the workplace, we need to make sure that the concept of anxiety is relevant for each cultural group and that it is responsive to the manipulations of our independent variable. Some cultural groups may not have a concept of anxiety, which then makes it difficult to operationalize this concept. In other culture groups, the concept of anxiety may exist but may be purely dispositional (e.g. a person is born anxious) and therefore be considered not responsive to situational variations.
- 3 In the psychometric and methodological literature, both the terms equivalence and invariance are used but they emerged from different philosophical research traditions. Measurement invariance is used widely in the psychometric literature and it is conceptually tied to the assumption that there are latent variables that 'cause' changes in behavioural indicators. This usage is tightly coupled to the use of structural equation modelling in psychometrics. Equivalence stems from an older literature which builds on classic discussions of scaling and does not

assume latent variables, but rather focuses on the comparability of scores in relation to scaling properties. Despite their different philosophical orientations, both approaches use the same set of statistical tests and draw similar conclusions (Fontaine, 2005). For clarification, functional equivalence only exists in the equivalence (scaling) literature, but not in the invariance (latent variable) literature.

## REFERENCES

- Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology, 71*(2), 230–244. doi:10.1037/0022-3514.71.2.230
- Baumert, A.; Schmitt, M.; Perugini, M.; Johnson, W.; Blum, G.; Wrzus, C. (2017). Integrating personality structure, personality process, and personality development. *European Journal of Personality, 31*, 503–528. doi: 10.1002/per.2115
- Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature, 483*(7391), 531–533. doi:10.1038/483531a
- Boer, D., Hanke, K., & He, J. (2018). On detecting systematic measurement error in cross-cultural research: A review and critical reflection on equivalence and invariance tests. *Journal of Cross-Cultural Psychology, 49*(5), 713–734. doi:10.1177/0022022117749042
- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110*(2), 203–219. https://doi.org/10.1037/0033-295X.110.2.203
- Brandt, M. J., Ilzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., ... van 't Veer, A. (2014). The Replication Recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology, 50*, 217–224. doi:10.1016/j.jesp.2013.10.005
- Cuddy, A. J. C., Fiske, S. T., Kwan, V. S. Y., Glick, P., Demoulin, S., Leyens, J.-P., ... Ziegler, R. (2009). Stereotype content model across cultures: Towards universal similarities and some differences. *British Journal of Social Psychology, 48*(1), 1–33. doi:10.1348/014466608X314935
- Dar-Nimrod, I., & Heine, S. J. (2011). Genetic essentialism: On the deceptive determinism of DNA. *Psychological Bulletin, 137*(5), 800–818. doi:10.1037/a0021860
- Doyen, S., Klein, O., Pichon, C.-L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE, 7*(1), e29081. doi: 10.1371/journal.pone.0029081

- Everett, J. A. C., & Earp, B. D. (2015). A tragedy of the (academic) commons: Interpreting the replication crisis in psychology as a social dilemma for early-career researchers. *Frontiers in Psychology, 6*, 1152. doi:10.3389/fpsyg.2015.01152
- Falk, A., & Heckman, J. J. (2009). Lab experiments are a major source of knowledge in the social sciences. *Science, 326*(5952), 535–538. doi:10.1126/science.1168244
- Farh, J.-L., Zhong, C.-B., & Organ, D. W. (2004). Organizational citizenship behavior in the People's Republic of China. *Organization Science, 15*(2), 241–253. doi:10.1287/orsc.1030.0051
- Fell, C. B., & König, C. J. (2016). Cross-cultural differences in applicant faking on personality tests: A 43-nation study. *Applied Psychology, 65*(4), 671–717. doi:10.1111/apps.12078
- Ferguson, G. M., Boer, D., Fischer, R., Hanke, K., Ferreira, M. C., Gouveia, V. V., ... Zenger, M. (2016). 'Get Up, Stand Up, Stand Up for Your Rights!' The Jamaicanization of youth across 11 countries through reggae music? *Journal of Cross-Cultural Psychology, 47*(4), 581–604. doi:10.1177/0022022116632910
- Fischer, R. (2012). Intersubjective culture: Indeed intersubjective or yet another form of subjective assessment? *Swiss Journal of Psychology, 71*(1), 13–20. doi:10.1024/1421-0185/a000067
- Fischer, R., Callander, R., Reddish, P., & Bulbulia, J. (2013). How do rituals affect cooperation? An experimental field study comparing nine ritual types. *Human Nature, 24*(2), 115–125. doi:10.1007/s12110-013-9167-y
- Fischer, R., Ferreira, M. C., Assmar, E. M. L., Baris, G., Berberoglu, G., Dalyan, F., ... Boer, D. (2014). Organizational practices across cultures. *International Journal of Cross Cultural Management, 14*(1), 105–125. doi:10.1177/1470595813510644
- Fischer, R., Ferreira, M. C., Milfont, T., & Pilati, R. (2014). Culture of corruption? The effects of priming corruption images in a high corruption context. *Journal of Cross-Cultural Psychology, 45*(10), 1594–1605. doi:10.1177/0022022114548874
- Fischer, R., & Karl, J. A. (2019). A primer to (cross-cultural) multi-group invariance testing possibilities in R. *Frontiers in Psychology, 10*, 1507. doi.org/10.3389/fpsyg.2019.01507
- Fischer, R., & Poortinga, Y. H. (2018). Addressing methodological challenges in culture-comparative research. *Journal of Cross-Cultural Psychology, 49*(5), 691–712. doi:10.1177/0022022117738086
- Fiske, A. P. (2002). Using individualism and collectivism to compare cultures – A critique of the validity and measurement of the constructs: Comment on Oyserman et al. (2002). *Psychological Bulletin, 128*(1), 78–88. doi:10.1037/0033-2909.128.1.78
- Fontaine, J. (2005). Equivalence. In *Encyclopedia of Social Measurement* (Vol. 1, pp. 803–813). Elsevier Science, New York
- Fu, J. H.-Y., Zhang, Z.-X., Li, F., & Leung, Y. K. (2016). Opening the mind. *Journal of Cross-Cultural Psychology, 47*(10), 1361–1372. doi:10.1177/0022022116652729
- Hofstede, G. (1980). *Culture's Consequences: International differences in work-related values*. Beverly Hills, CA: Sage.
- Hubbard, R., Vetter, D. E., & Little, E. L. (1998). Replication in strategic management: Scientific testing for validity, generalizability, and usefulness. *Strategic Management Journal, 19*(3): 243–254. doi:10.1002/(SICI)1097-0266(199803)19:3<243::AID-SMJ951>3.0.CO;2-0
- Islam, G., & Zyphur, M. J. (2009). Rituals in organizations. *Group & Organization Management, 34*(1), 114–139. doi:10.1177/1059601108329717
- Karl, J. A., & Fischer, R. (2018). Rituals, repetitiveness and cognitive load. *Human Nature, 29*(4), 1–24. doi:10.1007/s12110-018-9325-3
- Krátký, J., McGraw, J. J., Xygalatas, D., Mitkidis, P., & Reddish, P. (2016). It depends who is watching you: 3-D agent cues increase fairness. *PLOS ONE, 11*(2), e0148845. doi:10.1371/journal.pone.0148845
- Li, H., Rau, P.-L. P., & Salvendy, G. (2014). The effect of mixed American-Chinese group composition on computer-mediated group decision making. *Human Factors and Ergonomics in Manufacturing & Service Industries, 24*(4), 428–443. doi:10.1002/hfm.20391
- Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does 'failure to replicate' really mean? *American Psychologist, 70*(6), 487–498. doi:10.1037/a0039400
- McCambridge, J., Witton, J., & Elbourne, D. R. (2014). Systematic review of the Hawthorne effect: New concepts are needed to study research participation effects. *Journal of Clinical Epidemiology, 67*(3), 267–277. doi:10.1016/j.jclinepi.2013.08.015
- Mogan, R., Fischer, R., & Bulbulia, J. A. (2017). To be in synchrony or not? A meta-analysis of synchrony's effects on behavior, perception, cognition and affect. *Journal of Experimental Social Psychology, 17*, 13–20. doi:10.1016/j.jesp.2017.03.009
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. doi:10.1126/science.aac4716
- Oyserman, D., & Lee, S. W. S. (2008). Does culture influence what and how we think? Effects of priming individualism and collectivism. *Psychological Bulletin, 134*(2), 311–342. doi:10.1037/0033-2909.134.2.311
- Oyserman, D., Sorensen, N., Reber, R., & Chen, S. X. (2009). Connecting and separating mind-sets:

- Culture as situated cognition. *Journal of Personality and Social Psychology*, 97(2), 217–235. doi:10.1037/a0015850
- Pekerti, A. A., & Thomas, D. C. (2015). The role of self-concept in cross-cultural communication. *International Journal of Cross Cultural Management*, 15(2), 167–193. doi:10.1177/1470595814564767
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3), 638–641. doi:10.1037/0033-2909.86.3.638
- Singelis, T. M. (1994). The measurement of independent and interdependent self-construals. *Personality and Social Psychology Bulletin*, 20(5), 580–591. doi:10.1177/0146167294205014
- Smith, P. B., Fischer, R., Vignoles, V. L., & Bond, M. H. (2013). *Understanding social psychology across cultures: Engaging with others in a changing world*. London: Sage.
- Tsang, E. W. K., & Kwan, K.-M. (1999). Replication and theory development in organizational science: A critical realist perspective. *Academy of Management Review*, 24(4), 759. doi:10.2307/259353
- van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Thousand Oaks, CA: Sage.