

## **Apples and Oranges? Establishing Equivalence in Comparative Sport Policy Research**

Mathew Dowling<sup>1</sup> & Spencer Harris<sup>2</sup>

<sup>1</sup> Cambridge Centre for Sport and Exercise Sciences, Anglia Ruskin University, United

Kingdom

<sup>2</sup> College of Business, University of Colorado, United States

Submitted: September 2021

### **Abstract**

The notion of equivalence is important in the context of comparative studies, such as those that compare performance across sporting nations or those that compare good governance across different sport organisations. However, despite its importance, the concept has been interpreted and employed in different ways, resulting in the term being misunderstood or conflated. This article examines the concept of equivalence, discusses how issues of non-equivalence can arise, and identifies potential strategies that can be employed by researchers to ensure that is more appropriately addressed. We examine and apply three main types of equivalence (construct, sample and functional equivalence) to two empirical cases, (1) the SPLISS analysis of critical success factors in elite sport policy and (2) Play the Game's Sport Governance Observer to demonstrate how researchers attempt to overcome or at least mitigate the problems of equivalence and how, despite these efforts, there remain equivalence-related problems that limit the reliability and credibility of comparative elements of the study. We conclude our paper by discussing the implications for comparative sport research and specifically how future comparative sport research may be enhanced.

*Keywords:* equivalence, comparative, methodology, elite sport, governance

## Apples and Oranges? Establishing Equivalence in Comparative Sport Policy Research

Equivalence is an important concept for all comparative researchers. It is concerned with the extent to which the elements of the phenomenon under study can be considered comparable across different settings. Fundamentally, equivalence is concerned with sameness, ensuring that the social entities and instruments that form the basis of the comparison in one context are the same when compared to a different context, so that the similarities and differences observed are not simply a consequence of examining fundamentally different things (Hantrais, 2009; Jowell, 1998; Landman & Carvalho, 2017; Øyen, 1990). Within sport, this often means selecting countries with similar features or sport organisations that adopt similar roles in order to make intra-country comparisons or inter-organisational comparisons respectively. However, before we can examine how these issues can be understood and applied to sport in more detail, it is necessary to contextualise the discussion that follows with a general consideration of equivalence and the comparative approach in general.

To begin to understand the issue of equivalence and the comparative method generally, it is useful to reflect upon the common idiom that *it is not possible to compare apples with oranges*. This phrase gets to the core logic of comparative inquiry in highlighting the potential incommensurability of comparing two items that are typically not thought to be comparable. Hofstede (1998) challenges this assumption by arguing that although it may not be possible to compare apples with oranges as these are different objects, it is possible to compare them under the general category of fruits. Hofstede (1998) posits that if we examine apples and oranges as fruits (a fruitology), then it is possible to compare them on the basis of availability, price, colour, vitamin content and so on. Comparative analysis, then, is fundamentally about identifying both the similarities and differences across different social units. The assumption made by those that seek to make comparisons is that these social units are sufficiently similar

so as to permit meaningful comparisons, but at the same time they are also sufficiently diverse so as to reveal differences.

What can also be drawn from the above discussion is that it is necessary for researchers to infer and generalise in order to make meaningful comparisons. Inferences are a fundamental part of the scientific process and can be understood as “an attempt to infer beyond the immediate data to something broader that is not directly observed” (Della Porta, 2008, p. 199). Comparative researchers employ a range of concepts and methodological apparatus in order to make inferences between social units. A central question that underpins comparative analysis is how do we know that the use of concepts and application of methods in one context is the same in another? In other words, to what extent are the concepts and methodological apparatus equivalent? If they are not equivalent then we are not comparing like-for-like social units.

The aforementioned fruit analogy also serves to highlight the extent to which equivalence remains an important and longstanding issue within the general comparative methodological literature (Øyen, 1990; Dogan & Pelassy, 1990; Davidov *et al.* 2014; Ebbinghaus, 2005; Johnson, 1998; Landman & Carvalho, 2017; Mullen, 1995; Przeworski & Teune, 1966; Sartori, 1970) and, by extension, the comparative sport management and policy domain. As we will demonstrate, many of the methodological and practical issues faced by comparative sport policy researchers stem from issues regarding equivalence. Despite this, there has been limited explicit discussion of the philosophical assumptions or methodological issues related to comparative inquiry in sport in general (for exceptions see De Bosscher, 2018; Dowling & Harris, 2021; Dowling, Legg, Brown & Grix, 2018) including more specific, albeit pivotal, issues related to equivalence. A potential consequence of this has been a continued lack of conceptual clarity relating to equivalence and how it is employed by comparative researchers in sport.

To respond to this shortcoming, this article seeks to discuss the concept of equivalence, focusing on the three main types of equivalence - construct, sample and functional equivalence – in order to identify the key problems associated with ensuring equivalence and the potential strategies to overcome (or at least mitigate) equivalence related issues. To support the discussion, we draw upon two specific comparative sport studies: De Bosscher et al's (2006, 2015) Sport Policy Factors Leading to International Sporting Success (SPLISS and SLISS 2.0) and Play the Game's National Sport Governance Observer (Gearart, 2018<sup>1</sup>). We conclude by offering potential strategies to more appropriately address the concept of equivalence in future comparative sport studies.

### **Equivalence: Employing Non-Equivalent Terminology**

The notion of equivalence addresses an integral component of the comparative exercise that has been confused in its analysis and inconsistent in its application. Johnson (1998: 2) reinforces this view, stating that: 'perhaps in no field of inquiry...has this seemingly elementary concept been assigned as many alternative meanings and disaggregated into as many components as in the field of cross-cultural research.' Johnson (1998) goes onto identify no less than 52 *ways* in which the concept of equivalence has been employed across a number of disciplines including anthropology, business, sociology and political science. Although there is no universally agreed upon definitions between types of equivalence, there does appear to be some broad congruence regarding the essence and nature of the concept (cf. Øyen, 1990, Davidov et al., 2014; Hantrais, 2009; Johnson, 1998; Landman & Carvalho, 2017; Mullen, 1995; Przeworski & Teune, 1966; Stegmueller, 2011).

The conceptual coherence of equivalence is further complicated by the philosophical foundations underpinning research. Here, the researcher's ontological and epistemological

---

<sup>1</sup> To clarify, Play the Game developed the Sport Governance Observer concept, Gearart executed the study and authored the report.

position strongly influences the extent to which detailed attention is given to equivalence and the potential strategies that could be used to enhance it. Broadly speaking, research that is underpinned by an objectivist ontology and positivist epistemology tends to adopt a *nomothetic* approach that focuses on generating theories and concepts that can be applied universally (for example, constructs relating to self determination theory in the field of sport psychology and sport coaching). In contrast, research guided by a constructivist ontology and an interpretivist epistemology is more likely to take an *ideographic* approach emphasising the specificity of the case and the peculiarities of context and cultural specificity that do not permit overly simplistic or like-for-like comparisons (for example, qualitative comparisons of mass sport participation policy across different nations) (Hofstede, 1998; Przeworski & Teune, 1966). It is important to recognise these fundamental philosophical differences from the outset as they influence both the nature and extent to which researchers view equivalence to be an important issue and the type of strategies that may be used to overcome any potential problems.

Landman and Carvalho (2017) provide a useful distinction here in being able to address the relationship between research philosophy and comparative inquiry and the issue of equivalence. They identify three broad positions: universalist, relativist and the middle-position. The universalist viewpoint, aligning with objectivism/positivism, emphasises that the concepts and variables employed in comparative analysis must be able to travel analytically and have universal applicability. Relativists, aligning with constructivism/interpretivism, argue that all meaning is locally constructed and determined and therefore any attempt to make comparative claims across nations is exceptionally limited, even impossible. The middle ground position posits that the instruments employed can be adapted to be more culturally sensitive to facilitate equivalent and meaningful comparisons. We adopt this latter viewpoint (i.e. the middle ground) insofar as we believe that both concepts and variables – and therefore issues of equivalency – can be applicable across disciplines and domains but also recognise that knowledge is socially

constructed and context specific. In this sense, we believe that there are general lessons and examples of good practice that can be drawn from previous studies and the comparative methodological literature but at the same time we recognise that what constitutes ‘best practice’ and how (if at all) equivalency can be achieved depends on the nature of the research project and the researcher’s personal research philosophy. It is for this reason that our discussion of equivalence is deliberately tentative. The next section delineates some of the most common types of equivalence before illustrating how they have been addressed through empirical cases from the sport policy and management domain.

### **Construct, Sampling and Functional Equivalence**

#### **Construct Equivalence**

Construct equivalence is about ensuring that the concepts and instruments measure the same variables across different social units. In this sense, the notion of construct equivalence is defined here in a similar way to what Johnson (1998) describes more generally as *interpretive equivalence* i.e., the need to ensure that there is similarity of meaning between concepts. If the main aim of comparative analysis is to search for similarities and differences between nations, it is important to deploy concepts and instruments that measure equivalent variables across all cases. The essence of construct equivalence is captured in a rhetorical question posed by Przeworski and Teune (1966), ‘how can valid comparisons be made in cross-national research when so many terms and concepts differ in their meanings from country to country?’ (p. 551). For Przeworski and Teune (1966), ‘the critical problem in cross-national research is that of identifying equivalent phenomena and analysing the relationships between them in an equivalent fashion’ (p. 553). The general notion of ensuring equivalent concepts and the specific question posed by Przeworski and Teune (1966) are frequent and understandable challenges for comparative researchers as the social units and variables of one context are not always equal in another.

According to Johnson (1998), 'a measure can be identified as having [construct equivalence] to the degree that it exhibits a consistent theoretically-driven pattern of relationships with other variables across the cultural groups being examined' (p. 9). A commonly identified issue relating to construct equivalence within the comparative literature is language (Øyen, 1990; Jowell, 1998; Przeworski & Teune, 1966). This is also referred to as *translation equivalence* whereby the same items measure the same constructs across different cases (Mullen, 1995). For example, despite the vast array of different languages, it is often taken-for-granted that words are used equivalently across all cases. However, there are many words in certain languages that have different meanings in different contexts or have no equivalent at all. Additionally, the non-equivalence of language is not isolated to different languages. Construct equivalence issues can be evident between even the closest of matched languages. For example, many comparative studies erroneously assume that the English language is used consistently across English-speaking nations such as the US, Canada, and Australia for example. One only has to consider the numerous idiosyncrasies between US-English and British-English to appreciate the potential limitations of this approach. The consequence of erroneously assuming native English-speaking equivalence is that we rely on non-equivalent concepts which may adversely affect the reliability or credibility of the research. Importantly, the idiosyncrasies of non-equivalent language between US and British English might seem anecdotal, however, it serves to illustrate a critical equivalence issue insofar as 'different languages are not just equivalent means of defining and communicating the same ideas and concepts. In many respects they reflect different thought processes, institutional frameworks, and underlying values' (Jowell, 1998, p. 170). In order to overcome the problem of construct equivalence comparative researchers might need to go beyond the identification of a one-size-fits all construct to find similar, matched, like-for-like or equivalent indicators (Przeworski & Teune, 1966). Another potential strategy to overcome the issue of

non-equivalence of concepts is to employ a broader set of indicators or measurements rather than using single items (Przeworski & Teune, 1966). This ensures that any variance between units can be appropriately attributed to the independent variable in question rather than a misunderstanding or misappropriation of a non-equivalent concept. On this point, Przeworski and Teune (1966: 556) argue that ‘although complete equivalence is probably never possible, attempts can be made to measure equivalence if they are based on a set of indicators or observations.’

### **Sample Equivalence**

In addition to ensuring that concepts are equivalent, comparative researchers also need to check that their samples are equivalent (Øyen, 1990; Ebbinghaus, 2005; Hantrais, 2009; Jowell, 1998; Schuster, 2007). Sample equivalence refers to the degree to which the participants in the two (or more) cases being studied are representative of the entire target population under study and the extent to which the populations and samples are consistent in nature across cases (Neuliep, 2017). There are two important interrelated points regarding sample equivalence. The first concerns the general sample approach adopted. The second point relates to how these general approaches lead to more specific assumptions about the selected sample. In terms of general approaches to sampling, comparative researchers often distinguish between Most Similar System Designs (MSSD) and Most Different System Designs (MDSD) (Ancker, 2008; Przeworski & Teune, 1970). The MSSD approach is primarily based upon John Stuart Mills’s (1843) *method of difference* which states that

if an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance save one in common, that one occurring only in the former; the circumstance in which alone the two instances differ, is the effect, or cause, or a necessary part of the cause, of the phenomenon (p. 455).

In applying this logic of inference, those utilising the MSSD approach select cases that have as many similar features as possible with exception of the key factor or the variable that the researcher seeks to explore. The assumption here is that if a researcher controls as many confounding extraneous variables as possible then they are more likely to be isolate which factors might be causing a particular outcome. The MDSD, in contrast, is based upon Mill's (1843) *method of agreement* which states that "if two or more instances of the phenomenon under investigation have only one circumstance in common, the circumstance in which alone all the instances agree, is the cause (or effect) of the given phenomenon" (p. 454). Unlike the MSSD approach, the MDSD approach involves the deliberately selects cases that have different features and controls for potentially confounding variables in order to identify the key factor that is common across all cases.

The assumption of those that adopt a MSSD approach is that common system features are 'controlled for' and therefore any differences observed can be attributed to some explanatory variable. Equally, the MDSD approach assumes that all dependent variables do not change and are held constant. In practice, both approaches have equivalence related challenges. For example, the MSSD, it may not be possible to identify enough 'like-for-like' or similar cases and it may be possible to control for all confounding variables. This is commonly referred to as the '*too many variables not enough cases*' scenario (Ebbinghaus, 2005; Landman & Carvalho, 2017) whereby there are not enough cases to examine the key factors or explanatory variables. For the MDSD approach, it is often difficult to find cases that differ across all variables except the dependent variable and this typically can only be applied at the sub-system level (Przeworski & Teune, 1970).

The methodological decision regarding sample size, and by extension sample equivalence, also depends on whether the researcher adopts a variable-oriented or case-oriented approach (Ragin, 2014). Here, variable-oriented studies tend to focus on large-N, nomothetic,

quantitative approaches whereas case-oriented studies tend to utilise small-N, ideographic, qualitative approaches (see Table 1). The variable-oriented approach focuses on the relationship between variables and involves an extensive examination of social phenomena following the typical rules of statistical inference by controlling for any confounding variables and randomising the sample to avoid selection bias. In contrast, the case-oriented approach involves an intensive examination of a select number of units, often through purposeful selection and the application of qualitative techniques (see Ragin, 2014 for a detailed discussion on this point).

\*\*\*insert table 1 (overview of large-N and small-N sampling approaches) about here\*\*\*

These differences are important in regards to equivalence as they lead to more specific assumptions about the sample and sampling procedure. For example, it is often erroneously assumed that the selection of cases for variable-oriented approaches is randomised when in reality it is stratified. Ebbinghaus (2005: 136) describes this as the ‘illusion of random sampling’. In addition to this both variable and case-oriented approaches assume that it is at least theoretically possible to select equivalent units from a sample population. That said, the two approaches differ insofar as variable-oriented studies treat the sample as being the same (i.e. homogenous) and seek a randomised selection from the entire possible sample that could be selected (sample population) whereas case-oriented approaches assume that cases are heterogenous in nature and therefore select cases from the sample population in a deliberate and purposeful manner. In practice, both approaches face significant practical difficulties in being able to achieve either an entirely random or purposefully selected sample from any given population. The selection of units for comparative inquiry is rarely entirely random with cases often selected on the basis of practicality, feasibility or for historical reasons (Ebbinghaus,

2005). However, there are many ways guard against the issue of sampling equivalence. These include clearly articulating your philosophical position and general orientation including consciously thinking about and being aware of the assumptions relating to sampling, providing clear inclusion/exclusion criteria for sample selection, and explicitly recognising the limitations of the sampling approach adopted. Researchers can also avoid the ‘too many variables not enough countries’ scenario by ensuring that the sample selected is appropriate for the number of variables employed.

### **Functional Equivalence**

In addition to addressing construct and sample equivalence, comparative researchers should also ensure that the data collected and analysed has functional or measurement equivalence (Øyen, 1990, 2004; Davidov et al., 2014; Dogan & Pelassy, 1990; Ebbinghaus, 2005; Hantrais, 2009; Johnson, 1998; Jowell, 1998; Mullen, 1995; Schuster, 2007; Stegmüller, 2011; Schuster, 2007; Landman & Carvalho, 2017). Functional equivalence is centrally concerned with ensuring equivalence in the methods, measurement and procedures that underpin and guide the research process. Functional equivalence centres on fundamental issues about whether the variables identified and measures employed are standardised between cases throughout the research. The focus here is on the apparatus and the data used and whether they should be used for comparison. The essence of this issue is that just because data could be used for comparative purposes does not mean it should and that ‘comparable data are not necessarily usable data, but neither are usable data necessarily comparable’ (Schuster, 2007, p. 99).

The comparative literature reveals two major functional equivalence issues. The first relates to the identification of functionally equivalent units or institutions. A common assumption made by researchers is that similar entities with similar names are equivalent. This

is not always the case as the same entities can perform different functions and different entities can perform the same functions. On this point, Dogan and Pelassy (1990) note ‘the same performance may be accomplished in various countries by different [entities] and similar or comparable institutions may fulfil, in various countries, different tasks’ (p. 37). The key point here is that different structures within cases might perform the same function and equally the same structures can perform different functions. In both cases, there can be a potential lack of functional equivalence. This issue can be further illustrated by considering the roles, responsibilities and structures of the government in different nations. Within the UK, these entities refer to the people with authority to govern, the elected members who are given portfolios of responsibility through departments (e.g., the Department of Education or the Department of Digital, Culture, Media and Sport), and the elected representatives (members of Parliament) from different constituencies around the country. In some countries, however, these functions are performed by different or a combination of entities. For example, the US Congress performs many of the functions of Parliament, such as holding government to account and legislating, but clearly these units are not functionally equivalent.

The second, more specific functional equivalence issue relates to the deployment of instruments, measures, and procedures that are used to ensure equivalent responses. Some scholars have termed this *measurement equivalence* (Mullen, 1995). This seeks to address the question of whether the same measurements and apparatus hold across different units. Issues regarding this sub-category of functional equivalence relate to whether the questions employed, responses provided and data collected and analysed have functional equivalence (Dogan & Pelassy, 1990; Johnson, 1998; Mullen, 1995; Øyen, 1990). In regard to the questions employed, it is worth considering the obvious examples of units or measurements that are commonly different across national boundaries, for example Celsius in one country is Fahrenheit in another, miles versus kilometres per hour, dollars vs. euros, and so on. In much the same way

as you might decide to exchange currency or buy a power converter before you go on holiday, the questions within the instruments used to compare should also be converted to the measurements of the country in question. Hence why it is sometimes referred to as *calibration equivalence* because it involves the calibration of measurements in order to enable meaningful comparisons (Mullen, 1995).

A second functional equivalence issue relates to the whether the responses to the questions employed are comparable across cases. Mullen (1995) identifies two potential threats in relation to this sub-type of equivalence when employing survey instruments: familiarity with scaling and scoring methods and response bias. In terms of familiarity with scaling and scoring, some countries may not be as familiar with these types of research methodological approaches which may pose a potential threat to reliability. Also, responses may differ between countries by virtue of differing social norms and national cultures. For example, some countries may have populations who are normatively more assertive, confident or outspoken whilst the normative values of other nations may place greater importance on politeness or humility. The outcome of these differences, when aggregated across given populations, may result in vastly different responses to the questions posed. These issues might be spotted through the identification of inconsistencies in responses or tendencies towards the mean (which are particularly prevalent when respondents are uncertain). In addition, researchers can check that there is no systematic bias in response to questions.

A third functional equivalence issue relates to whether the data collected and analysed has functional equivalence. This is particularly the case when only a small number of questions are employed within a questionnaire, for example, to encapsulate a particular variable under investigation. To resolve this issue, multiple indicators can be used to measure or capture each variable. This ensures that the variance in response from the data collected is actually reflective of the observed differences. The issue of data collection equivalence also relates to more

practical considerations of what primary or secondary data is available for making comparisons. For primary data, this is often a feasibility and data access issue and a trade-off between cost and resources and the ability to include countries. For secondary data, this issue relates to whether the data that exists is appropriate for comparisons and, if it is, whether or not it should be used for comparative inquiry.

The following part of this paper draws upon two empirical case studies from the comparative sport policy and management domain to further illustrate how issues of equivalence can arise and how researchers have attempted overcome, or at least mitigate, them. Through these cases, we aim to demonstrate the utility of the above distinctions for better understanding the nature and extent of equivalence issues in comparative sport research as well as identify and elaborate on some potential strategies that can be employed to enhance equivalence for future comparative research in sport.

### **Empirical Case Studies: SPLISS and the Sport Governance Observer**

This section draws upon two large-scale comparative studies that have been recently conducted within the sport policy and management domain: De Bosscher and colleagues' (2006, 2015) *Sport Policy Factors Leading to International Sporting Success* (SPLISS and SPLISS 2.0 studies) and Play the Game's *National Sport Governance Observer* (SGO) Gearart's (2018). These cases were selected primarily on the basis that they are both well-known within the sport management and policy literature, but also due to the present authors' previous involvement with collecting data for these studies. Our discussion of equivalency below is therefore simultaneously informed by both our independent interpretation and judgement of the research as well as our own experiences as co-investigators within them. In particular, many of the issues of equivalence identified below stem from our own experiences and the challenges and difficulties we faced when attempting to apply these frameworks and their modus operandi to

collect data within our own contexts. We recognise, therefore, that the nature and extent of these issues may have varied considerably between context to context and from researcher to researcher. Nonetheless, we have attempted to focus on what we considered to be the most substantive equivalent related issues within both projects. We begin by providing a brief overview of these studies before examining how they attempted to address issues of equivalence.

### **Sport Policy Factors Leading to International Sporting Success (SPLISS)**

In 2006, De Bosscher and colleagues developed a conceptual model highlighting a number of sport policy factors that lead to international sporting success (SPLISS) (DeBosscher et al., 2006). SPLISS phase 1.0 was empirically tested in 2006/7 by comparing sport policy factors across six nations (Belgium, Canada, Italy, the Netherlands, Norway and the United Kingdom). The overall purpose of the pilot study was to model the relationship between sport policy factors and international success. A total of nine sport policy factors form the basis of the study (financial support for sport, organisation and structure of sport policies, foundation and mass sport participation, talent identification and development, athlete career support, training facilities, coaching provision and coach development, international competition, and scientific research and innovation) with performance being assessed through the analysis of 103 critical success factors distributed across the nine pillars.

Methodologically, the SPLISS research is ostensibly based on a mixed methods approach involving an elite sport policy inventory and an elite sport climate survey (De Bosscher et al., 2006). The policy inventory utilised data from interviews with policy agents combined with analysis of key policy documentation. The elite sport climate survey measured the primary users (e.g. athletes, coaches, performance directors) perceptions of success in each nation by responding to a number of dichotomous (yes/no) and/or ordinal (five-point Likert

scale) questions on a total of 103 critical success factors. The Likert scale responses were calculated by multiplying the response values resulting in scores where 1.00=highly developed, 0.75=sufficiently developed, 0.50= reasonably developed, 0.25=insufficiently developed and 0=not developed. Following this initial scoring, the sub-factor scores were totalled for each critical success factor and then aggregated into a total percentage score for each of the nine pillars. These scores were then used to develop a traffic light scale (red, amber, green) depicting the relative performance of each nation against the 103 critical success factors together with a radar chart showing the nation's performance across the nine pillars.

In 2015, SPLISS phase 2.0 was launched and involved 15 nations (De Bosscher et al., 2015). This follow up study sought to better understand the effectiveness and efficiency of sport policies and their relation to elite sport success. The same conceptual model was used based on the nine pillars with an adapted 96 critical success factors and an additional 750 sub factors relating to the 96 critical success factors. The output from SPLISS 2.0 was similar to SPLISS 1.0 with the addition of statistical analysis (z-scores, distance from mean, cumulative probability scores) in order to sharpen insights about the elite sport systems and the extent to which policy factors influence elite sport success. See De Bosscher et al. (2006, 2009, 2015) for a more detailed summary of their methodological approach and the distinctions between SPLISS 1.0 and 2.0.

### **The National Sport Governance Observer (NSGO)**

The NSGO was developed by Geeraert (2018) and supported by Play the Game's<sup>2</sup> ongoing efforts to drive improvements in the governance of Olympic sport organisations. In short, the NSGO provides an evaluation framework to assess good governance in National Sport

---

<sup>2</sup> Play the Game is an initiative run by the [Danish Institute for Sports Studies](https://www.playthegame.org) (Idan), aiming at raising the ethical standards of sport and promoting democracy, transparency and freedom of expression in world sport. See here for further information: <https://www.playthegame.org>

Federations. The principal aim of the NSGO is ‘to assist and inspire federations to enhance the quality of their governance by measuring governance and building capacity’ (Geeraert, 2018, p. 11). Importantly, the NSGO project sought to support good governance by providing a consistent framework to allow for meaningful inter-country (across countries) and intra-country (across federations) comparisons (Geeraert, 2018). The design of the NSGO was informed by a detailed study of good governance and the identification of a checklist of elements considered to be essential to the good governance of sport organisations (Geeraert, 2015). This study highlighted transparency, democratic processes, internal accountability, and societal responsibility as four core dimensions of good governance. The NSGO operationalises each of these dimensions through a total of 46 principles which are measured in specific terms via a total of 274 dichotomous indicators. The framework includes a detailed *modus operandi*, explaining the standardised data gathering process; providing definitions for the dimensions, principles, and indicators; detailing the key evaluation criteria for each indicator; and providing a clear system of scoring for each indicator. On this latter point, a score of 1 is recorded if evidence demonstrates that the requirements of the indicator are met, and a score of 0 is given if the requirements are not met. There are a total of seven indicators in the transparency dimension, 13 in the democratic processes dimension, 14 indicators for internal accountability, and 12 for societal responsibility. The scoring for each dimension is calculated as a percentage score. The overall score for each federation is calculated by aggregating all four dimensions to get a percentage score with all indicators, principles and dimensions equally weighted. See Geeraert (2018) for a more detailed overview of the methodological approach.

### **Ensuring Equivalence through SPLISS and the NSGO**

The following section analyses the issues relating to construct, sample and functional equivalence in applying the SPLISS framework and protocol to compare elite sport success

across sporting nations and the NSGO in comparing performance across NSFs and differing national contexts.

***Construct equivalence.*** In both cases, we assumed that language would be the major problem or limitation when applying frameworks developed in Europe to contexts outside of Europe. However, this was not the case. For the SPLISS studies, the issue of language appears to have been adequately managed by clearly defining key terms and concepts, translating the framework and survey tools into multiple languages (survey instruments were translated into 5 languages in SPLISS 1.0 and 12 languages in the SPLISS 2.0 study), and by utilizing local researchers to execute the study (SPLISS 1.0 involved 9 researchers across 6 nations and SPLISS 2.0 involved 58 researchers across 15 nations). These attempts to mitigate equivalence issues, however, create new challenges. For example, the use of local researchers brings issues concerning inter-observer and study reliability. Here, despite having work protocols in place, local researchers have discretion to assess their own nation's sporting system, with limited oversight or checks and balances, thus permitting considerable risk for researcher bias and the suppression of cultural specificity. Furthermore, these individual researchers (or small groups of researchers) not only come with their own research interests and philosophical backgrounds, which may impact what they view as important and how data is reported, but they are also effectively responsible for making observations and collecting data to determine the effectiveness of an entire nations' sporting system.

For the NSGO study, the major challenge was seen to be in the detail of the 274 indicators and the way in which these indicators aligned to the normative 'thought processes, institutional frameworks, and underlying values' of the European context with no explicit regard for other national contexts (Jowell, 1998, p. 170). This leads to a situation where many of the indicators are at best tenuously related or entirely inappropriate to contexts outside

Europe. For example, a group of indicators make reference to the NSF's multi-annual policy plan. However, not only do US-based NSFs not produce such a plan, but such processes are not an accepted part of good governance practices or routines in the US context. Similarly, a series of indicators referred to the NSF's annual report. However, no US-based NSF produces an annual report in the same way that may be seen in Europe. Instead, each NSF is legally required to make publicly available its annual 990 filing to the Internal Revenue Service. The 990 filing should not be considered equivalent to the annual report as it fails to represent the breadth of a typical annual report and, more importantly, it is a legal requirement from the Internal Revenue Service rather than a proxy of good governance. Additionally, another set of the indicators throughout the framework make reference to the NSF's General Assembly but no US-based NSF had a General Assembly and only the larger sports had the capacity to offer an annual conference, congress or meeting of members. On reflection, General Assemblies (or their equivalent) were considered an unusual practice in the US rather than one representing good governance. Similar issues of construct equivalence were evident within the SPLISS studies, for instance the critical success factors used often employed constructs which aligned with a European-centric model of sport including whether countries had a centralised governing agency (pillar 2), a nationally co-ordinate facility database (pillar 6) or involvement of military in supporting elite athletes (pillar 5) amongst others.

Importantly, Gearart (2018) indirectly attempts to address the problems of construct equivalence in two distinct ways. First, the standardised scoring mechanism of NSGO allows the researcher to enter *not applicable* rather than a score of 1 or 0 in the case that a federation cannot reasonably be expected to comply with an indicator. However, in such cases the opportunity for the NSF to score points and thus aggregate an overall higher good governance score is lost. The effect is then further exacerbated when there are a large number of indicators that are deemed not applicable, as was the case in our evaluation of US NSFs of sport. This

ultimately could be viewed as a subtle form of ethnocentrism (Dogan & Pelassy, 1990) in that the assumption being made here is that different approaches must be indicative of poor governance. Second, the NSGO framework is multi-dimensional insofar as there are a large number of indicators used to examine each principle. In total, 42 indicators are used for transparency, 55 for democracy, 90 for accountability and 87 for societal responsibility. SPLISS 1.0 adopted 105 critical success factors and SPLISS 2.0 utilised 96 critical success factors and 750 sub-factors across the nine pillars. While such an approach may mirror Przeworski and Teune's (1966) call to use a broader set of indicators rather than using single items, if the multiple indicators simply make continued reference to the same set of processes or activities that do not exist (i.e. multi-annual policy plan, annual report or general assembly), then the original rationale for using multiple indicators is fundamentally flawed. The inherent danger of continuing to add multiple indicators to a study, much in the way that SPLISS and NSGO have done, is the 'too many variables, not enough cases' scenario which produces indeterminant findings (Ebbinghaus, 2004; Landman & Carvalho, 2017). See Henry et al. (2020) for a detailed discussion relating to the limitations of variable-oriented approaches.

***Sample equivalence.*** Despite explicit inclusion criteria being developed for the SPLISS studies, the sample for these studies is primarily based on pragmatism, in particular, the willingness and ability of nations (or their respective NOCs or elite sport agencies) to take part in the study. In SPLISS 2.0, 'any nation interested was invited to participate under the condition that they were able to collect the comprehensive data set and follow the research protocol' (De Bosscher et al., 2015, p. 66). There also appeared to be many researchers from other nations that were interested in taking part in the SPLISS and SPLISS 2.0 studies but did not have sufficient resources to do so. The decision to adopt a pragmatic approach to sampling, whilst methodologically convenient, may result in selection bias as the sample does not include

appropriate representation from nations at either end of the success continuum, with limited representation from nations who dominate the medals (the UK took part in SPLISS 1.0 but not SPLISS 2.0, the U.S., China and Russia did not take part in SPLISS 1.0 or SPLISS 2.0) or nations that do not typically win any medals. Consequently, the sample provides a stratified sample providing a partial picture based on data from self-selecting, resource-rich researchers without due consideration of the importance of accessing data from the most and least successful nations.

In contrast to the SPLISS study, there are two major considerations in the sample selection for the NSGO study, the countries included and the specific sports selected in each country. The first issue (i.e., country selection), is a challenging one as it requires that a local researcher/research team is willing and able to conduct the research in a specific country. While the project coordinator can be proactive and solicit certain nations, the commitment and resources to initiate and complete the study in each nation must come from the researcher/research team who is/are usually living and working in the country, and therefore have some level of familiarity with sport governance related policies, structures and practices.

The majority of nations for the phase 1.0 NSGO project were from Europe, primarily as the project was funded by European Union and Council of Europe. Brazil and Montenegro were added to the phase 1.0 project as associate external partners. The six countries involved in the phase 2.0 NSGO project responded to a call for research participants. Thus, the general approach to the sample of nations can be best described as responsive (responding to those who are willing and able to participate), pragmatic (making sure each country has the interest and resources to complete the study), and variable-oriented (insofar as the variables under study are represented by the dimensions, principles and indicators of good governance). However, unlike other variable-oriented approaches, the NSGO theoretically addresses the problem of sampling equivalence and the assumption of homogeneity across cases by preparing country

reports. The country reports set out the broader socio-cultural and sporting context and how this context may impact the good governance of NSFs. Thus, when the NSGO scores are read as part of each country chapter the comparative reliability is enhanced by giving explicit attention to this context and acknowledging how it shapes good governance and influences the similarities and differences in good governance when compared to other countries. Furthermore, the study also acknowledges that when reporting quantitative data, the complexity of context and the detail of cases is often lost to the sharp and the visually engaging glare of the numerical data. More practically, the problem could be better addressed by providing a detailed overarching analysis of the comparative data, both qualitatively and quantitatively, which clearly addresses the key similarities and differences, between nations and NSFs, and how these similarities and differences are influenced by sport and non-sport issues.

The second consideration, mentioned above, involves the selection of sports. While all nations involved in the NSGO phase 1.0 and 2.0 studies were permitted to select sports of their choice, all nations were required to focus on five sports including football (soccer), tennis, handball, swimming, and athletics, ‘to allow for the cross-country comparisons of the governance of the same sports’ (Geeraert, 2018, p. 23). While such an approach goes some way to ensuring consistency, it is important to avoid the assumption that the same entities (e.g. sport specific NSFs across different nations) perform the exactly the same function across nations. Similarly, it is equally important to recognise the extreme diversity in the structure and resources of NSFs representing the same sport across different nations, a point that we return to in more detail in the following section.

***Functional equivalence.*** The SPLISS 2.0 study took steps to minimise the problems of functional equivalence of SPLISS 1.0 insofar as it triangulated sources of data to avoid

overreliance on single sources and prepared detailed work protocols to ensure consistency of approach across the 15 different nations involved in the study. For example, SPLISS researchers used a sport policy inventory to support their semi-structured interviews with senior officials together with an elite sport climate survey completed by athletes, coaches and performance directors. However, despite these developments, there remain significant functional equivalence problems with how comparisons are made within the study. In relation to inputs and financial support, for example, it is difficult to ensure consistency across nations as the definitions of expenditure and what is included and excluded as financial support in one national context is unlikely to be consistent across other national contexts. De Bosscher and colleagues recognise this by stating that “transnational comparisons of sport expenditure are challenging as expenditure definitions and sport delivery mechanisms vary considerably from nation to nation” (De Bosscher et al., 2015, p. 109). This comment not only reveals the complexity of the presentation of budgets, budget headings and the decisions on what monies get allocated to allocated to what areas but also the different levels of investment into sport (national, regional and local) and the variety of sources that may invest in sport in differing national contexts. Consequently, the SPLISS researchers pursued a pragmatic approach by focusing in expenditure at a national level by government, lotteries, and nationally coordinated sponsorship only. In doing so, they do not account for local government or private sector funding as the “data is not available in a format that permits transnational comparison” (p. 109). However, while pragmatic, this approach is clearly limited in its ability to capture a true and full picture of sport expenditure, thus reinforcing the reality of the nature of the specific practical challenges that impact comparative research.

Another example of problematic functional equivalence is the use of participation data where no universally agreed upon or standardised protocol or datasets exists for comparing sport participation across countries. De Bosscher et al. (2015) also acknowledge this in their

study by stating that “comparing sport participation internationally is a tremendously difficult task, because of the different standards of defining sport and determining frequency and intensity” (p. 181). De Bosscher and colleagues attempt to respond to this challenge by drawing upon multiple indicators (and not just sport participation data) including physical education opportunities within school, teacher certification, and availability and co-ordination of extra-curricular competitions. Furthermore, rather than relying on national participation survey data, the SPLISS studies utilised the European barometer (EB) survey (for SPLISS 1.0 and 2.0) in addition to the International Social Survey Programme (ISSP) (for SPLISS 2.0) as an indication of general sport participation levels. Not only are these non-sport specific surveys, but they have been challenged as useful means of comparison in their own right (Höpner & Jurczyk, 2015). The use of these surveys also prevents comparison of mass sport participation in four of the 15 countries taking part in the SPLISS 2.0 study. While De Bosscher and colleagues found no significant correlation between participation and elite sport outcomes they also make clear that this “might be a problem of measurement (comparable data)” and that “if we had compared nations with huge differences in sport participation we might have found different results” (De Bosscher et al., 2015, p. 184). This latter point also lends further support to the viewpoint that country selection was stratified. These specific examples of expenditure and participation from the SPLISS studies not only raise questions about the extent to which other variables might explain success (i.e. potential confounding variables) but also reveal more fundamental issues regarding the degree to which these data can be appropriately standardised and contextualized to allow for meaningful comparisons.

Both the SPLISS and the NSGO studies adopted detailed *modus operandi* and definitions in an attempt to ensure functional and construct equivalence. The NSGO were guided by a detailed framework accompanied by exact procedures explaining how researchers should select their sample, gather their data, and analyse their results. For example, researchers

were required to follow a standardised process involving sample selection, data gathering and analysis involving six phases (with detailed instructions for each phase): (i) sport organisations are selected and contacted, (ii) initial data is collected and initial scoring completed, (iii) the initial data is shared with the NSF and the NSF is given the opportunity to provide feedback and additional evidence, (iv) the data is re-evaluated and a second preliminary scoring of the indicators is completed, (v) the secondary evaluation is shared with the NSF and final feedback and any additional evidence is submitted, and (vi) the final evaluation and scoring is completed and submitted to the NSF. Once this process has been completed for all NSFs in one country, the data is submitted to Play the Game who undertake a check of the data to minimise inconsistencies and errors.

However, beyond the surface level, there are procedural issues that are more problematic in nature. First, the extent to which sport specific NSFs can be considered functionally equivalent units across nations is questionable due to national variations in the governance, structure and demand for and supply of sport as well as variations in the size and resources of sport-specific NSFs across nations. This same issue is relevant to the SPLISS studies insofar as they focus on National Olympic Committees, when, in fact, some countries rely more heavily on other quasi/non-governmental agencies (than the NOC) to support and invest into elite sport. As addressed in the sampling equivalent section, this latter issue is important as the diversity of organisation types may challenge the relevance and applicability of generalised good governance principles. Furthermore, organisational diversity is particularly important in the context of the NSGO because the NSFs' ability to cooperate was heavily influenced by the size of the NSF (with larger NSFs typically being more willing and able to cooperate than smaller NSFs), and NSFs cooperation shared a significant positive correlation with higher good governance scores. Put another way, size does matter as larger NSFs are able

to commit resources to the project and provide the researcher with evidence that they meet the requirements of the good governance indicators.

The NSGO also reflects problems concerning *measurement* equivalence insofar as the framework measures a range of items that do not reflect the national model of sport. Here, we argue that a number of indicators contained within the societal responsibility dimension of the framework lean heavily toward the European model of sport as evidenced in the attention given to indicators that address sport for development (indicators 39.1-39.6) or sport for all (indicators 45.1-45.6). This would not be a problem if other national models of sport followed the European model, but this is not the case. For example, the US-based NSFs tend to be focused on performance development and revenue generation with limited capacity (or concern) to address sport for development or sport for all. Thus, to measure their performance against these issues and to make assertions about good governance is misleading. Of course, it is possible to identify and score these items as *not applicable* but again such an approach has an overall negative effect on the NSF's good governance score.

In a related yet contrasting point, functional equivalence is problematic when NSFs score positively in good governance terms for practices where the NSF is simply following legal requirements. In the US context, this is the case for indicators 15.1-15.4 where the Amateur Sport Act requires that NSFs must have 20% of athlete representation on NSF boards and other indicators (specifically 37.1-37.12) where the SafeSport Act states that federations must follow set policies and procedures relating to cases involving accusations of sexual abuse in sport.

In summary, the application of the SPLISS framework to compare elite sport success across sporting nations and the NSGO in comparing performance across NSFs and differing national contexts demonstrates how all three equivalence related issues may arise within a comparative study. The cases also reveal that many of these issues occur due to the underlying

assumptions which underpin the application of a framework or instrument from one context to another. What can also be drawn from the cases is the clear overlap between different types of equivalence. Issues of construct equivalence and ensuring that the concepts utilised are comparable across cases clearly has implications for functional equivalence and vice versa, thus demonstrating the complexities of how these issues occur in practice. Finally, the cases also identify a number of potential strategies that could be employed to overcome (or at least mitigate) equivalence-related issues. These include the drafting modus operandi with definitions, employing multiple indicators, utilising local researchers, country-specific reports, standardised selection of sports, and selecting a broad range of sport organisations. These potential strategies along with those identified previously are summarised in Table 2.

\*\*\*insert Table 2 (potential strategies for enhancing equivalence) about here\*\*\*

### **Conclusion and Implications**

This article examined equivalence, discussed how issues of (non-) equivalence can be problematic for comparative research, and identified potential strategies that can be employed by comparative sport researchers to ensure better equivalence – summarised in Table 2. In so doing, three main types of equivalence (construct, sample and functional equivalence) have been delineated along with a consideration of how (if at all) sport comparative sport researchers have attempted to overcome (or at least mitigate) these issues by drawing upon two empirical examples from the elite sport policy and management research domain.

In considering the implications of our discussion for comparative inquiry in general, the concept of equivalence can be understood in much a similar fashion to how researchers ensure research quality within their studies by systematically addressing the validity and reliability of their projects. For comparativists, the issue of equivalence is an additional layer of complexity

that needs to be addressed within in the research design and considered throughout the research process when attempting to carry out social inquiry. Ensuring equivalence is fundamental as without it studies run the risk of misrepresentation.

What can also collectively be drawn from the above discussion, and the comparative literature in general, is the need to build-in strategies to ensure better equivalence throughout the research process. In our view, it is not possible to achieve what Verba (1978) described as *complete equivalence* – the achievement of total—i.e. complete construct, sample and functional—equivalence. For Dogan and Pelassy (1990: p. 16), what comparative researchers ‘should seek is not paralyzing perfection, but the most satisfying approximation to it’. For this reason, it is argued that the best that comparative sport researchers can hope to achieve is something as close as possible to *complete equivalence* by attempt to make *imperfect* comparisons of approximations of social reality (Øyen, 1990).

Another general consideration it is that although some of the potential strategies identified above may be useful, they could understandably be perceived as being quite difficult or practically unrealistic to incorporate into a comparative research design. It is argued here – as it has been suggested elsewhere – that although equivalence is of central concern to comparative researchers, ‘not all forms of equivalence are necessarily created equal’ (Johnson, 1998, p. 30). Quite to what extent that one type of equivalence is more important than another is beyond the scope of this paper. Nonetheless, it is appropriate to recognise that it is not realistic or practical for those who are seeking to make comparisons to employ all of the abovementioned strategies.

Finally, even if researchers were somehow able to overcome or at least mitigate some of the aforementioned issues of equivalence, it is equally important to recognise that ‘comparative studies, will always be defeated to some extent by differences between nations in matters or taxonomy and technique. Indeed, their very starting point is that important differences exist

between nations in their behaviours, circumstances and attitudes' (Jowell, 1998, p. 173). Jowell's remarks suggest that issues of equivalence are ultimately unavoidable. After all, comparative studies are entirely predicated upon the basis that there are different characteristics or features that makes them unique and worth comparing. From this viewpoint, the issues of equivalence are likely to be a constant feature of comparative inquiry and a perennial concern for those attempting to make comparisons. For this reason, we hope that our discussion of equivalence supports the need to move from a universalist to a middle-ground approach to comparative analysis whereby sport policy/management researchers recognise the unique limitations and challenges of conducting comparative analysis compared to other forms of social inquiry.

### References

- Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology*, 40(1), 55–75. <https://doi.org/10.1146/annurev-soc-071913-043137>
- De Bosscher, V. (2018). A mixed methods approach to compare elite sport policies of nations. A critical reflection on the use of composite indicators in the SPLISS study. *Sport in Society*, 21(2), 331-355. 10.1080/17430437.2016.1179729
- De Bosscher, V., De Knop, P., Van Bottenburg, M., & Shibli, S. (2006). A conceptual framework for analysing sports policy factors leading to international sporting success. *European Sport Management Quarterly*, 6(2), 185–215. 10.1080/16184740600955087
- De Bosscher, V., Shibli, S., Westerbeek, H., & Van Bottenburg, M. (2015). *Successful elite sport policies: an international comparison of the sports policy factors leading to international sporting success (SPLISS 2.0) in 15 nations*. Meyer & Meyer Sport (UK).
- Della Porta, D. (2008). Comparative analysis: case-oriented versus variable-oriented research. In D. Della Porta & M. Keating (Eds.), *Approaches and methodologies in the social sciences: a pluralist perspective* (pp. 198–222). Cambridge University Press.
- Dogan, M., & Pelassy, D. (1990). *How to Compare Nations: Strategies in Comparative Politics*. Chatham House.
- Dowling, M., & Harris, S. (2021). *Comparing Sporting Nations: Theory and Method*. Meyer & Meyer Sport.
- Dowling, M., Brown, P., Legg, D., & Grix, J. (2018). Deconstructing comparative sport policy analysis: assumptions, challenges, and new directions. *International journal of sport policy and politics*, 10(4), 687-704.

- Ebbinghaus, B. (2005). When less is more: selection problems in large-n and small-n cross-national comparisons. *International Sociology*, 20(2), 133–152. 10.1177/0268580905052366
- Geeraert, A. (2018). National Sports Governance Observer: Final report. Copenhagen: Play the Game.
- Hantrais, L. (2009). *International Comparative Research: Theory, Methods and Practice*. Palgrave Macmillan UK.
- Henry, I., Dowling, M., Ko, L. M., & Brown, P. (2020). Challenging the new orthodoxy: a critique of SPLISS and variable-oriented approaches to comparing sporting nations. *European Sport Management Quarterly*, 20(4), 520-536.
- Hofstede, G. (1998). A case for comparing apples with oranges: International differences in values. *International Journal of Comparative Sociology*, 39(1).
- Höpner, M., & Jurczyk, B. (2015). How the Eurobarometer Blurs the Line between Research and Propaganda, MPIfG Discussion Paper 15/6.
- Johnson, T. P. (1998). Approaches to equivalence in cross-cultural and cross-national survey research. *ZUMA-Nachrichten Spezial*, 3, 1–40.
- Mullen, M. (2016). Diagnosing measurement equivalence in cross-national research. *Journal of International Business Research*, 26(3), 573–596.
- Jowell, R. (1998). How comparative is comparative research? *American Behavioral Scientist*, 42(2), 168–177. 10.1177/0002764298042002004
- Landman, T., & Carvalho, E. (2017). *Issues and Methods in Comparative Politics* (4th ed.). Routledge.
- Lijphart, A. (1971). Comparative Politics and the Comparative Method. *American Political Science Review*, 65(3), 682–93.

- Mills, M., van de Bunt, G., & de Bruijn, J. (2006). Comparative research: persistent problems and promising solutions. *International Sociology*, 21(5), 619–631. 10.1177/0268580906067833
- Øyen, E. (1990). *Comparative Methodology: Theory and Practice in International Social Research*. SAGE Publications.
- Øyen, E. (2004). Living with imperfect comparisons. In: P. Kennet (Ed.). *A Handbook of Comparative Social Policy*. (pp. 276-293). Elgar Publishing
- Przeworski, A., & Teune, H. (1966). Equivalence in cross-national research. *The Public Opinion Quarterly*, 30(4), 551–568.
- Przeworski, A., & Teune, H. (1970). *The Logic of Comparative Social Inquiry*. Wiley-Interscience.
- Ragin, (2014). *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies* (2nd ed). University of California Press.
- Sartori, G. (1970). Concept misinformation in comparative politics. *American Political Science Review*, 64(4), 1033–1053. 10.2307/1958356
- Schuster, J. M. (2007). Participation studies and cross-national comparison: proliferation, prudence, and possibility. *Cultural Trends*, 16(2), 99–196. 10.1080/09548960701299815
- Stegmueller, D. (2011). Apples and Oranges? The problem of equivalence in comparative research. *Political Analysis*, 19(4), 471–487. <https://doi.org/10.1093/pan/mpr028>
- Verba, S. (1978). *Participation and Political Equality: A Seven-Nation Comparison*. University of Chicago Press.

**Table 1: Overview of Large-N/Small-N sampling approaches**

	Large-N	Small-N
Number of cases	>20	5-19
Sample selection	Random/stratified	Purposeful
Emphasis	Extensive	Intensive
Theory	Theory building and testing	Theory building
Approach	Variable focused	Case-based
Analysis	Quantitative	Qualitative
Strengths	Internal validity	External validation
Weaknesses	<ul style="list-style-type: none"> <li>- Finding cases</li> <li>- Non-random sample</li> <li>- Problem of contingency</li> </ul>	<ul style="list-style-type: none"> <li>- Finding cases</li> <li>- Problem of contingency</li> </ul>

**Table 2: Potential strategies for enhancing equivalence**

<b>Construct equivalence</b>	<b>Sample equivalence</b>	<b>Functional equivalence</b>
Define all concepts and key terms and share with all researchers through an overarching project guidance document/modus operandi	Clearly specify the inclusion and exclusion criteria used for sampling purposes	Use researchers with local knowledge of the national context and implement checks and balances to minimise problems of bias
Develop and apply multiple indicators to guard against overreliance on limited data	Reflect upon and explicitly state the sampling approach guiding the study (MSSD/MDSD)	Where possible and appropriate, use standardised data-sets
Translate all research instruments and associated resources into all languages relevant to the study	Make sure that the sampling approach is proportional to the variables employed	Standardise quantitative data utilising statistical techniques (if not standardised)
Ensure that all instruments and resources are translated using consistent protocols and procedures	Utilise randomisation techniques for large-N studies	Use empirical techniques to validate data, where appropriate (for example, confirmatory factor analysis)
Develop a pilot project to empirically test the equivalence of concepts	Be aware of problem of selection bias (i.e. choosing cases to support positive outcome) and generate sample that best reflects the field of study	Use mixed methods, multiple sources with an emphasises on primary (rather than secondary) data
	Be genuine and explicit about the limitations of the study	Undertake triangulation of data to confirm, reject or to further discuss findings
		Check for inconsistent responses between cases as this may indicate an issue of functional equivalence