OECD Education Working Papers No. 107

Report on Social Desirability, Midpoint and Extreme Responding in TALIS 2013

Fons J. R. Van de Vijver,

Jia He

https://dx.doi.org/10.1787/5jxswcfwt76h-en

OECD

Organisation de Coopération et de Développement Économiques
Organisation for Economic Co-operation and Development

19-Nov-2014

English - Or. English

DIRECTORATE FOR EDUCATION AND SKILLS

EDU/WKP(2014)6
Unclassified

Cancels & replaces the same document of 12 November 2014

REPORT ON SOCIAL DESIRABILITY, MIDPOINT AND EXTREME RESPONDING IN TALIS 2013

OECD Education Working Paper No. 107

This paper was prepared by Fons J. R. Van de Vijver and Jia He and was funded and coordinated by the American Institutes for Research, USA.

The authors can be contacted at: Department of Culture Studies, Tilburg University, P.O Box 90153, 5000 LE Tilburg, The Netherlands, Phone: +31 13 4662528, Email: fons@fonsvandevijver.org and jamis.he@gmail.com.

The Board of Participating Countries (BPC) to TALIS (Teaching and Learning International Survey) provided valuable feedback on a draft version of this paper.

This working paper has been authorised by Andreas Schleicher, Director of the Directorate for Education and Skills, OECD.

Julie Belanger, Analyst, TALIS
Email: julie.belanger@oecd.org
Tel: +(33-1) 45 24 91 93

JT03366557

English - Or. English

OECD EDUCATION WORKING PAPERS SERIES

Working Papers describe preliminary results or research in progress by the author(s) and are published to stimulate discussion on a broad range of issues on which the OECD works. Comments on Working Papers are welcome, and may be sent to the Directorate for Education and Skills, OECD, 2 rue André-Pascal, 75775 Paris Cedex 16, France.

Comment on the series is welcome, and should be sent to edu.contact@oecd.org.

This working paper has been authorised by Andreas Schleicher, Director of the Directorate for Education and Skills, OECD.

---

www.oecd.org/edu/workingpapers

ABSTRACT

This paper investigated the effects of response styles in cross-cultural contexts. The authors used data on social desirability, extreme and midpoint responding, and the scale scores of 17 constructs related to the teaching profession, which were collected from 76,887 teachers in 18 countries in the Teaching and Learning International Survey (TALIS). Main findings are: (1) a 10-item social desirability scale demonstrated partial invariance of a positive and a negative impression management factor; (2) a general response style, representing a continuum ranging from response amplification to moderation, was extracted with social desirability and extreme responding as positive indicators and midpoint responding as a negative indicator; (3) social desirability and the general response style at the country level were negatively correlated with country affluence and educational achievement; (4) social desirability and the general response style were more strongly correlated with constructs of teacher efficacy and job satisfaction than other constructs at both the individual and country level; and (5) correction of response styles had negligible effects on cross-cultural differences in these constructs.

RÉSUMÉ

Ce document de travail étudie les effets des styles de réponses au sein des contextes interculturels. Les auteurs se sont servis des données en matière de « désirabilité sociale », des réponses extrêmes et médianes, et des scores des échelles de mesures de 17 constructions relatives à la profession d'enseignant. Ces données ont été recueillies auprès de 76,887 enseignants dans 18 pays de l'enquête internationale sur l'enseignement et l'apprentissage (TALIS). Les principales conclusions sont : (1) l'échelle de mesure de 10 items en matière de désirabilité sociale a démontré une invariance partielle du facteur de gestion du processus d'influence des perceptions positives ou négatives. (2) une réponse d'ordre général, représentant un continuum s'étendant d'une amplification de la réponse à la modération, a été extraite avec la désirabilité sociale et une réponse extrême comme indicateurs positifs et réponse médiane comme indicateur négatif ; (3) la désirabilité sociale et la réponse d'ordre général au niveau des pays étaient liées de manière négative avec la richesse des pays et le niveau d'éducation ; (4) la désirabilité sociale et la réponse d'ordre général étaient plus fortement liées avec les constructions relatives à l'efficacité des enseignants et la satisfaction dans l'exercice de leur profession que les autres constructions au niveau individuel et au niveau des pays ; et (5) la correction des styles de réponses avaient des effets négligeables sur les différences interculturelles de ces constructions.

TABLE OF CONTENTS

Tables

Figures

Boxes

PART 1: INTRODUCTION

1.1 Preamble

We report the results of a study of response styles (social desirability, midpoint and extreme responding) in the 2013 Teaching and Learning International Survey (TALIS), coordinated by the OECD. The study set out to examine to what extent country differences in core TALIS constructs, such as teacher job satisfaction and self-efficacy, are influenced by these response styles and, if so, whether statistical corrections for such styles would affect the nature and size of the country differences in core constructs. Before discussing these study aims in more detail in section 1.5, we describe the background of the study.

When survey participants are asked to indicate to what extent they endorse statements about views, attitudes, values, or other psychological constructs, they can, either implicitly or explicitly, portray themselves in a particular, often favourable way. This is known as impression management in surveys. Social desirability, one of the most frequently studied impression management strategies, refers to the tendency of respondents to reply in a manner that will be viewed favourably by others (Paulhus, 1991).

In addition to social desirability, three other impression management strategies have been frequently studied in the context of surveys: acquiescent, extreme, and midpoint responding. Acquiescence means that participants show a tendency to endorse items irrespective of the item contents. Extreme responding means that participants show a tendency to opt for the extremes of the response scale, favouring response options like "strongly disagree" or "strongly agree". Midpoint responding refers to participants' tendency to overuse the middle category in odd-numbered Likert response scales or the middle categories in even-numbered scales (Yang, Harkness, Chin, & Villar, 2010). These impression management strategies are collectively known as response styles. This study focuses primarily on social desirability and extreme and midpoint response styles.

Response styles have been studied in psychology since the 1950s. Six decades of research have yielded much empirical data; however, conceptual progress has been very modest. Traditionally, response styles are considered nuisance factors that threaten the validity of findings. Recently a new view on response styles was proffered according to which these styles are culturally preferred communication styles that are embedded in cultural values and personality. In a nutshell, there is still no widely accepted theory of response styles including social desirability. Not surprisingly, different measures of response styles have been proposed and are used side-by-side, although their results do not always converge. Acquiescent, extreme, and midpoint responding are usually measured indirectly from items of various substantive constructs, whereas social desirability is measured with independent scales. The scale that is used in this study, (an adapted version of) the Marlowe-Crowne Social Desirability Scale (Crowne & Marlowe, 1960), has been administered most frequently. There is less agreement about how acquiescent, extreme, and midpoint responding should be measured (as described in more detail below).

1.2 Response Styles in the TALIS Project: Some History

The Teaching and Learning International Survey (TALIS) is an OECD study that includes a principal survey and a teacher survey on various aspects of their working conditions and the learning environments across countries. The first round of the TALIS study was conducted in 2008 and the second round in 2013.

A recurrent issue in the analysis of the first round of the TALIS data involved the role of response styles. Notably during meetings of the groups responsible for the questionnaire development and analysis, participants repeatedly discussed to what extent the country differences that were observed in the core constructs of the TALIS project would be susceptible to response styles. The underlying question was: to

what extent can we trust country differences and constructs like teacher self-efficacy and satisfaction with profession as reflecting valid differences; or, to what extent are these differences influenced by response styles? Data on social desirability were not available; however, scores on acquiescence could be computed on the basis of the available survey data. Analysing data from the first round of TALIS, Vieluf, Kunter, and van de Vijver (2013) found that one of the strongest correlates of country differences in teacher self-efficacy was acquiescence. The authors proposed to further disentangle the effects of response styles on country differences in teacher self-efficacy and other constructs.

Based on interest among the participating countries to more carefully investigate the potential impact of response styles on measures in TALIS, the 2013 administration of the study included an optional set of questions to measure social desirability. Of the 34 participating countries and educational jurisdictions in 2013, 20 included the social desirability scale in the questionnaires administered to teachers. Of those 20 countries, 18 agreed to participate in an in-depth analysis of the responses of teachers to the social desirability questions. These 18 countries are Abu Dhabi (United Arab Emirates), Brazil, Chile, Croatia, Estonia, Finland, France, Iceland, Korea, Latvia, Malaysia, Mexico, Poland, Portugal, Serbia, the Slovak Republic, Spain, and the United States.

1.3 Models of Social Desirability

Studies of social desirability do not reveal the same number of factors, and different instruments yield very different factor structures. The Marlowe-Crowne Social Desirability Scale consists of 33 descriptions of "*highly desirable but rare*" and "*highly undesirable but common*" behaviours. It measures respondents' tendency to present themselves in a positive light (Crowne & Marlowe, 1960). Initially conceptualised as unidimensional (Crowne & Marlowe, 1964), this scale has been repeatedly found to be multidimensional. However, there is no complete convergence as to the number of factors underlying the scale (Barger, 2002; Loo & Loewen, 2004). In the present study we focus on the two-factor solution as the presumably most frequently reported number that yields interpretable factors. Examples of such studies can be found in the work by Millham (1974) and Ramanaiah, Schill, and Leung (1977). They found a two-dimensional structure of the Marlowe-Crowne Social Desirability Scale: *positive impression management* (the tendency to attribute socially desirable characteristics to oneself) and *negative impression management* (the tendency to deny socially undesirable characteristics), which is in line with the basic motives of self-presentation: looking good and avoiding to look bad (Schütz, 1998).

Social desirability can challenge the validity of psychological measures, as respondents with high scores on social desirability tend to respond according to how they think people in their immediate environment would like them to react, where the interpretation of scores usually assumes that responses are not contaminated by response styles. In this line of thinking, social desirability is a nuisance factor that should be minimised, either through a careful research design or statistical corrections (Nederhof, 1985). However, this prevailing view has been challenged. In an alternative interpretation, social desirability is considered part and parcel of the psychological makeup of individuals (Uziel, 2010) and reflects culturally preferred ways of communication, which are associated with various other cultural characteristics (van Hemert, van de Vijver, Poortinga, & Georgas, 2002). At the country level, social desirability has been found to be negatively associated with country affluence and individualism (Johnson & van de Vijver, 2003). Van Hemert et al. (2002) studied the *Lie Scale* (a measure of social desirability) from the *Eysenck Personality Questionnaire* in a cross-cultural meta-analysis. They confirmed the above mentioned associations, and reported a negative association of social desirability with autonomy from the *Schwartz Value Survey*.

1.4 Integration of Social Desirability and Other Response Styles to a General Factor

Social desirability can be considered an example of a response style. Other examples are acquiescent, extreme, and midpoint responding. What all these tendencies have in common is participants' systematic use of certain response anchors (tendencies) on some basis other than the target construct. In the case of social desirability, this involves the desire to create a positive impression and avoid a negative impression. The three other response styles distinguished here (acquiescent, extreme, and midpoint responding) are particularly relevant when Likert response scales are used. Acquiescence refers to the tendency to agree with the statement in the item, irrespective of the contents of the item (yea-saying). Midpoint and extreme responses refer to tendencies to either avoid or frequently use the extreme categories of a response scale.

On the basis of their definitions, correlations between the four response styles can be expected and, indeed, have been documented. Social desirability and extreme responding are positively associated; both are positively related to desirable personality traits such as extroversion and conscientiousness (Austin, Deary, & Egan, 2006; de Vries, Zettler, & Hilbig, 2013; Musek, 2007). Extreme responding is negatively correlated with midpoint responding, reflecting the contrast of being decisive versus evasive (Baumgartner & Steenkamp, 2006; Naemi, Beal, & Payne, 2009). Acquiescent responding has been found to be positively related to extreme responding, as it is often operationalised as a weak form of extreme responding (i.e., the endorsement of the positive end of a scale is taken as both acquiescent and extreme responding). In a stricter operationalisation where the positive end of the scale is not taken as an indicator of acquiescent responding, the meaning of acquiescent responding changes as it is then more closely related to midpoint responding.

Because different studies focus on different response styles and the operationalization of response style measures differ across studies, findings on response styles are difficult to replicate or generalise. We argue that the various styles share basic tendencies and have common elements that may underlie a general response style factor. Therefore, we argue that it is necessary to integrate different response styles (study the shared meaning in these response styles), which may help to create more consistency in findings. Using both (conventional) indirect and direct self-reports of response styles, He and van de Vijver (2013) confirmed that a general response style factor can integrate the four response styles. Social desirability and extreme responding are positive indicators and acquiescent (not including the extreme agreement category) and midpoint responding are negative indicators of the general response style factor. At the individual level, this general response style factor is positively related to desirable personality traits such as agreeableness, extroversion, conscientiousness, and openness, as well as well-being, individualistic values, and self-promotion and appraisal regulation. At the country level, this factor is related to values and personality traits that pertain to "fitting in" and the avoidance of ambiguity (He, van de Vijver, Domínguez, & Mui, in press).

1.5 Aims of the Present Report

The aims of the present report are fivefold. The first three aims are psychometric and refer to the measurement invariance and meaning of social desirability and the general response style factor:

1. We check if the two-factor structure of social desirability (positive impression management and negative impression management) is supported by the teacher data in TALIS 2013 and if the scale shows measurement invariance across countries.

2. Extending the study of social desirability to other related response styles, we examine whether social desirability, midpoint responding and extreme responding correlated in the expected direction and can be taken to constitute a general response style. That is, we expect a positive association between social desirability extreme responding, and negative associations of midpoint

responding with the other two styles. This integration of different response styles to a general factor is expected to help create consistency in findings.

3. We examine the psychological meaning of social desirability and the general response style by correlating them with other country characteristics.

4. The next two aims deal with the consequences of social desirability and the general response style factor for individual- and country-level differences in the core TALIS constructs:

5. We examine the effects of social desirability on the core constructs from TALIS 2013.

6. We examine the effects of the general response style factor on the core constructs from TALIS 2013.

PART 2: METHOD

2.1 Sample

We focus on the TALIS 2013 data from lower secondary school teachers (i.e., ISCED Level 2)[1] who had valid responses on the social desirability scale across the 18 countries that administered the social desirability scale items in the teacher questionnaire and gave permission for inclusion of their data in the analyses. In this section, we describe the data cleaning and final sample compositions.

2.1.1 Data Cleaning

There were four sources of missing data. First, 14 of the 34 TALIS countries (Australia, Flanders (Belgium), Bulgaria, Alberta (Canada), Cyprus[2,3], the Czech Republic, Denmark, Italy, Japan, the Netherlands, Norway, Singapore, Sweden, and England (United Kingdom) did not administer the scale. Second, Israel and Romania did not make their data available for analysis. Third, 8% of the teachers that did participate had incomplete, omitted or invalid responses. Fourth, less than 1% of teachers in each country were omitted from the analyses because their responses were deemed highly unlikely to represent a real attitude (see Table 1). These included teachers who, in their responses to the 10-item social desirability scale with response options ranging from 1 (*totally disagree*) to 7 (*totally agree*), provided a score of 7 on all items (see the list of items in Table 3); that is they responded "*totally agree*" to all items. There were no teachers with only score of 1, or "totally disagree," on any items.

Table 1 Number of Teachers with a Score of 7 on All Social Desirability Items

Country	Number of teachers with valid responses	Number of teachers who indicated a score of 7 on all social desirability scale items
Abu Dhabi (UAE)	2187	23
Brazil	12490	27
Chile	1460	10
Croatia	3478	5
Estonia	2981	3
Finland	5890	1
France	2911	0

1 The classification in the International Standard Classification of Education (ISCED 1997) of the UNESCO can be found at http://www.uis.unesco.org/Education/Pages/international-standard-classification-of-education.aspx

2 Note by Turkey: The information in this document with reference to "Cyprus" relates to the southern part of the Island. There is no single authority representing both Turkish and Greek Cypriot people on the Island. Turkey recognises the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Turkey shall preserve its position concerning the "Cyprus issue".

3 Note by all the European Union Member States of the OECD and the European Union: The Republic of Cyprus is recognised by all members of the United Nations with the exception of Turkey. The information in this document relates to the area under the effective control of the Government of the Republic of Cyprus.

Table 1 Number of Teachers with a Score of 7 on All Social Desirability Items *(continued)*

Country	Number of teachers with valid responses	Number of teachers who indicated a score of 7 on all social desirability scale items
Iceland	1176	1
Korea	2783	20
Latvia	4103	5
Malaysia	3839	9
Mexico	5015	7
Poland	3694	9
Portugal	6809	3
Serbia	3549	8
Slovak Rep.	3404	1
Spain	9214	8
United States	1904	0
Total	**76887**	**140**

2.1.2 Demographics of the sample

In total, there were 76 887 teachers in 18 countries with valid, complete responses on the social desirability scale. The average age (in years) of teachers ranged across countries from 38 (Abu Dhabi (UAE)) to 48 (Estonia). In most of the 18 countries, there were more female than male teachers. In all countries, teachers' highest level of educational attainment on average was ISCED Level 5A (i.e., Bachelor's degrees and Master's degrees from universities or equivalent institutions). Ninety-two percent of teachers across countries had this as their highest level of education. Among all the teachers, years of experience in the current school ranged from an average of 4 (Korea) to 15 (Latvia), and years of experience as a teacher in total ranged from an average of 13 (Abu Dhabi (UAE)) to 22 (Latvia). Age, percentage of females, and years of experience as a teacher at the current school and in total differed significantly across countries (Wilks' Lambda = .74, $F(85, 324748) = 249.31$, $p < .01$, $\eta^2 = .06$). The detailed information of the sample demographics is presented in Table 2.

Table 2 Sample Demographics

Country	Sample Size	Mean Age	Proportion of Females	Average Educational Attainment*	Experience as teacher - current school (years)	Experience as teacher - in total (years)
Abu Dhabi (UAE)	2187	38	.60	3	5	13
Brazil	12490	39	.67	3	7	14
Chile	1460	41	.62	3	10	15
Croatia	3478	42	.74	3	12	15
Estonia	2981	48	.83	3	14	21
Finland	5890	44	.70	3	10	15
France	2911	42	.66	3	9	16

Table 2 Sample Demographics *(continued)*

Country	Sample Size	Mean Age	Proportion of Females	Average Educational Attainment*	Experience as teacher - current school (years)	Experience as teacher - in total (years)
Iceland	1176	45	.70	3	10	15
Korea	2783	43	.70	3	4	17
Latvia	4103	47	.88	3	15	22
Malaysia	3839	39	.72	3	7	14
Mexico	5015	42	.50	3	11	16
Poland	3694	42	.75	3	11	17
Portugal	6809	45	.73	3	11	20
Serbia	3549	43	.66	3	11	14
Slovak	3404	43	.81	3	12	18
Spain	9214	46	.59	3	9	18
United States	1904	42	.67	3	8	14
Total	**76887**	**43**	**.69**	**3**	**10**	**17**

Note. * A score of 3 in educational attainment represents ISCED Level 5A which includes Bachelor's degrees and Master's degrees from universities or equivalent institutions

2.2 Instrument

2.2.1 Social desirability

Social desirability was measured with a short, adapted version of the Marlowe-Crowne Social Desirability scale. Ten items in the original Marlowe-Crowne Social Desirability Scale were selected on two criteria: (1) items should not have an ambiguous meaning and (2) items should be appropriate in different cultural contexts. The shortened scale was expected to maintain the construct validity and improve cross-cultural comparability. Given the focus of TALIS on the teacher workforce, item wording was adapted to a teaching and learning context. Five items were worded positively (positive impression management) and the other five items were worded negatively (negative impression management). The positive and negative impression management dimensions were expected to show a negative association. These items were translated into the languages of the respective countries; a rigorous translation verification process (which applied to all TALIS questionnaire items) was implemented to ensure consistency and comparability across the many contexts. More information on quality control procedures such as translation verification can be found in the TALIS 2013 Technical Report (OECD, 2014).

This 10-item scale was first tested in the Field Trial involving all 34 TALIS countries/economies: Abu Dhabi (UAE), Alberta (Canada), Australia, Brazil, Bulgaria, Chile, Croatia, Cyprus, the Czech Republic, Denmark, Estonia, Finland, Flanders (Belgium), France, Iceland, Israel, Italy, Japan, Korea, Latvia, Malaysia, Mexico, the Netherlands, Norway, Poland, Portugal, Romania, Serbia, Singapore, the Slovak Republic, Spain, Sweden, England (United Kingdom), and the United States. The structure of the two dimensions (i.e., positive and negative impression management) was supported, and the reliability (Cronbach's alpha) for the two dimensions was .65, and .57, respectively. The scale showed acceptable metric invariance, yet scalar invariance was not achieved. Based on feedback from participating countries and colleagues at the OECD, Item 5 was slightly reworded from "I am jealous of teachers who are successful" to "I feel threatened by teachers who are very successful". In the end, the following items presented in Table 3 were used in the main survey.

Table 3 The Social Desirability Items in the TALIS 2013 Main Survey

Item	Positive Impression Management	Negative Impression Management
1. I always listen carefully to students	X	
2. I am confident about my judgments about students	X	
3. I have doubts about my ability to succeed as a teacher		X
4. I have always been honest with myself about my teaching qualities	X	
5. I feel threatened by teachers who are very successful		X
6. I have said things that hurt colleagues' or students' feelings		X
7. I feel angry when colleagues express ideas different from my own		X
8. I help students and colleagues in trouble	X	
9. I admit when I do not know something if a student asks a question in class	X	
10. I am irritated by students who ask for favours		X

2.2.2 Indexes of extreme and midpoint responding

With the exception of the social desirability scale, all the other attitudinal items in the TALIS 2013 teacher survey had a 4-point response scale (the international version of the teacher questionnaire can be found at www.oecd.org/edu/school/Questionnaires%20TALIS%202013.pdf). We utilised these attitudinal items with 4-point "*strongly disagree*" to "*strongly agree*" response anchors to construct indexes of extreme and midpoint responding; it may be noted that even-numbered response scales such as the 4-point scale used in TALIS do not have a midpoint, in which case the responses adjacent to the midpoint (choices of 2 and 3) are taken as evidence for midpoint responding. We decided not to include acquiescence in the present study. The main reason is the incoherent operationalisation in the literature. Acquiescence is sometimes operationalised as a weak form of extreme responding, by taking *agree* and *strongly agree* as its indicators. It is our experience that acquiescence and extremity have a very strong positive correlation in such a conceptualisation, which limits the value of using both (He and van de Vijver, 2013). However, when acquiescence is operationalised as the proportion of *agree* responses (leaving out *strongly agree* responses), acquiescence is much more a measure of modesty as it then correlates negatively with extreme responding and positively with midpoint responding. We avoided this ambiguity by only using midpoint and extreme responding as indexes of response styles.

It has been recommended that at least 15 items of heterogeneous content should be used to derive valid and reliable response style indexes (De Beuckelaer, Weijters, & Rutten, 2010). Indexes of extreme and midpoint responding were extracted with non-overlapping items of 4-point Likert anchors from various items (e.g., other TALIS questionnaire items related to topics such as teacher feedback, personal beliefs in teaching, and school climate), in order to avoid data dependency between indexes. Specifically, fifteen items were randomly chosen to construct the extreme responding index. The average inter-item

correlation among the chosen items was .03, indicating sufficient heterogeneity in item content. The original responses were recoded as extremity endorsement (i.e., original respondent responses on Likert scales of 1 and 4 were recoded as 1) and extremity non-endorsement (i.e., original respondent responses of 2 and 3 were recoded as 0). The reliability of the 15 recoded items was .73. The extremity endorsement from the 15 items was then averaged as an indicator of extreme responding.

A similar procedure was employed for the midpoint responding index. Another non-overlapping 15 items that used a 4-point Likert style scale were randomly chosen (average inter-item correlation .02), and recoded as midpoint responding endorsement (i.e., original respondent responses of 2 and 3 were recoded as 1) and midpoint responding non-endorsement (i.e., original respondent responses of 1 and 4 were recoded as 0). The reliability of the 15 recoded items was 70. The average of the endorsement was taken as the index of midpoint responding.

PART 3: RESULTS

We describe the results in five parts (following the order of the study aims):

1. We report the measurement invariance of the two factors of social desirability: positive and negative impression management.

2. We expand the analysis of social desirability to a general response style, which is a combination of social desirability, extremity, and (reversed) midpoint responding.

3. We correlate the two social desirability factors and the general response style factor at country level with other country characteristics, including social and educational development indicators, to shed light on the psychological meaning of response styles.

4. We examine the impact of correcting for social desirability in the 17 core constructs measured in TALIS.

5. We extend the analysis of the fourth step to the general response style factor.

3.1 Measurement Invariance Analyses of the Social Desirability Scale

3.1.1 Model testing

Determining measurement invariance amounts to conducting a set of statistical tests to establish whether the same construct is measured across cultures and whether items behave in the same way across cultures. Measurement invariance is a prerequisite for the cross-cultural comparison of the scale scores. In order to compare scores across countries, it is necessary to show that there is a sound statistical basis for such comparisons. Invariance is addressed here using structural equation modeling. Multigroup confirmatory factor analyses were conducted to test invariance of the structure of the scale across countries. We checked configural invariance (i.e., the same indicators loaded on the same latent variables across countries), metric invariance (i.e., factor loadings on the latent variables were constrained to be equal across countries) and scalar invariance (i.e., items were constrained to have the same intercepts across countries). Invariant intercepts across countries indicate that cross-country differences in the means of the observed items are due to differences in the means of the underlying constructs, whereas differences in intercepts suggest some systematic differential attractiveness of an item across cultures even if persons from these cultures would score the same on other items. Scalar invariance is the prerequisite for valid mean comparison across countries. The model fit was evaluated by Chi-square tests, Comparative Fit Index (CFI: usually considered acceptable above .90), the Tucker Lewis Index (TLI: acceptable above .90), and the Root Mean Square Error of Approximation (RMSEA: acceptable below .06).

As metric and scalar invariance of all items are often difficult to find in studies involving many countries (Byrne & van de Vijver, 2010), we resort to partial measurement invariance. Partial measurement invariance means that at least one, but not every, parameter of the factor model (usually a factor loading or measurement intercept) is left free in at least one country. These country-specific parameters are usually indicated on theoretical or statistical grounds (the latter refer to a substantial

improvement in fit by lifting invariance constraints of parameters across countries, as measured by modification indexes).

Initial modification indices suggested to account for the similarity in meaning between Item 8 "I help students and colleagues in trouble" and Item 9 "I admit when I do not know something if a student asks a question in class", and between Item 3 "I have doubts about my ability to succeed as a teacher" and Item 5 "I feel threatened by teachers who are very successful", thus the error terms of the two pairs of items were correlated. The results of the invariance testing are presented in Table 4. Configural invariance and metric invariance were rather well supported (with fair fit statistics, given the large sample size and number of countries involved). Scalar invariance was not supported; values of incremental values of χ^2/df and CFI suggested that scalar invariance could not be accepted.

To improve the fit, partial measurement invariance was examined. In our case, the constraints on the loadings and intercepts of Item 9 and Item 10 were released because they showed the largest variations across countries. The fit of partial metric invariance improved slightly and that of partial scalar invariance improved significantly, compared with the full metric and scalar invariance models.

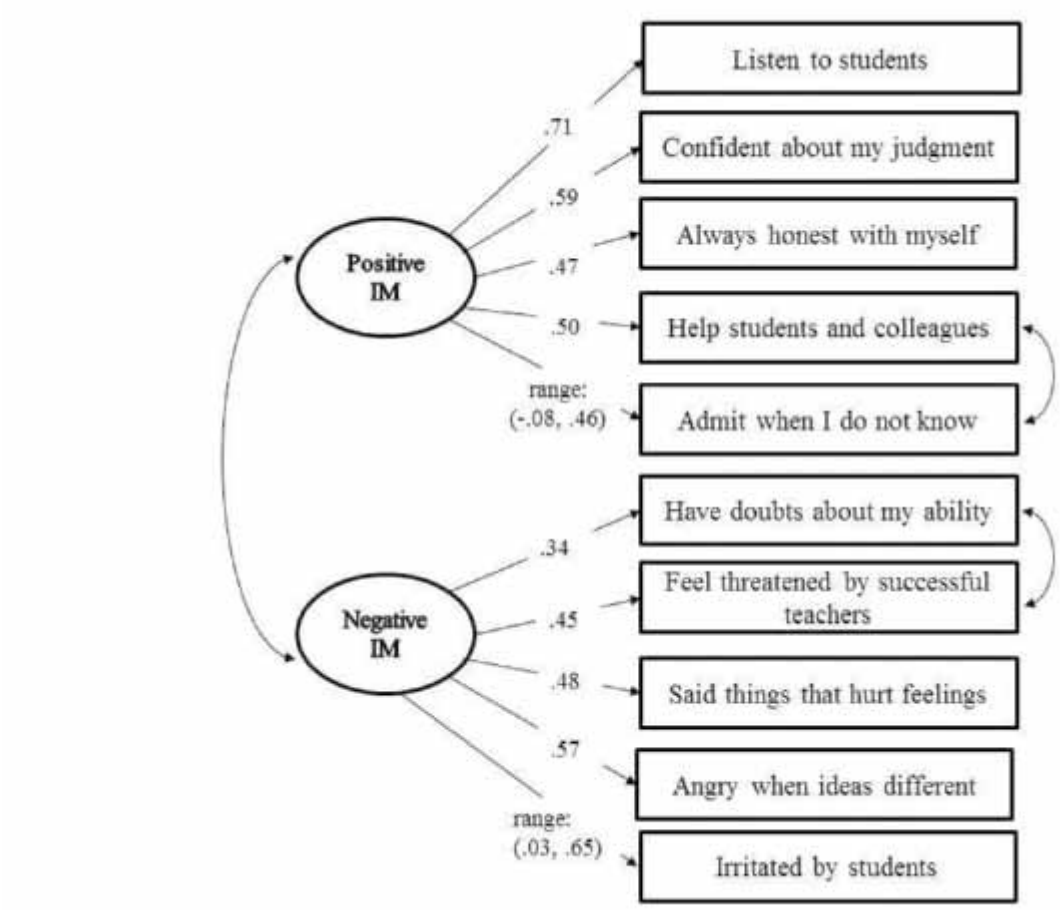
Table 4 Model Fit of the Multigroup Confirmatory Factor Analyses

Model	χ^2/df	CFI	TLI	RMSEA	$\Delta\chi^2/\Delta df$	ΔCFI	ΔRMSEA
Configural Invariance	10.462**	.935	.908	.011			
Metric Invariance	10.966**	.915	.903	.011	13.096**	-.020	.000
Scalar Invariance	76.979**	.196	.262	.031	353.459**	-.793	.020
Partial Metric Invariance	*10.408***	*.923*	*.909*	*.011*	*1.109***	*-.012*	*.000*
Partial Scalar Invariance	39.306**	.626	.628	.022	183.368**	-.309	-.011

Note: Most restrictive model with an adequate fit is printed in italics. CFI = Comparative Fit Index; TLI = Tucker–Lewis Index; RMSEA = Root Mean Square Error of Approximation. **p < .01.

Partial metric invariance was the most parsimonious model that fitted well. Standardised factor loadings (of the partial metric invariance solution) are presented in Figure 1. Loadings of all items were mostly significant and in the expected direction. The constraints on two items were released, therefore the loadings varied from country to country. Item 9 "I admit when I do not know something if a student asks a question in class" loaded positively on the positive impression management factor in most countries except Malaysia (range: -.08, 46), and Item 10 "I am irritated by students who ask for favours" loaded positively on the negative impression management factor in all countries (range: .03, .65). The correlation between the two latent factors on average was -.50 (range: -.65, -.40). The correlation between the error terms of Item 8 "help students and colleagues" and Item 9 "admit when I do not know" on average was .16 (range: -.01, .29), and that between Item 3 "have doubts about my ability" and Item 5 "feel threatened by successful teacher" was .14 (range: .05, .23). These correlations are small and would presumably not have implications for the interpretation of the model outcomes (or future applications of the instrument). The evidence for scalar invariance is very weak at best.

Figure 1 The Standardised Solution for the Social Desirability scale in the Partial Metric Invariance Model



3.1.2 Evaluating the effects of lack of invariance

Given the lack of support for scalar and partial scalar invariance of the social desirability model, there is no statistical ground for comparing the observed score means. However, a similar lack of scalar invariance for all scales was observed in both TALIS rounds. It has been argued that it is uncommon to find support for scalar invariance in studies involving huge samples in many countries, because the invariance tests are sensitive to sample size; it is not always clear whether the reasons for the lack of invariance are major misspecifications of the model or minor misspecifications that do not have many consequences for the rank order of the country means (Byrne & van de Vijver, 2010). Some studies have suggested that partial and full invariance models can sometimes produce very similar patterns of cross-cultural differences. To establish to what extent the size of the observed cross-cultural differences is affected by the lack of scalar invariance in the data, we examined the differences for the scalar invariance model and partial scalar invariance model with both a latent mean and an observed scale score approach (as detailed below).

First, with Brazil as the reference group (given that it had the largest sample size among all countries), the *latent means* of both factors (positive impression management and negative impression management) for each country were estimated in the scalar invariance or the partial scalar invariance model. The rankings and latent means per country are presented in Table 5. A higher score on positive impression management represents a stronger tendency to engage in positive impression management (i.e., endorsing the positively worded items), whereas a higher score on negative impression management represents a

stronger tendency to engage in negative impression management (i.e., endorsing the negatively worded items). Scores above (below) zero of a country indicate a stronger (weaker) tendency to display impression management than Brazil. For example, teachers in Korea had the weakest tendency to engage in positive impression management and the strongest tendency to engage in negative impression management.

Table 5 Country Rankings and Latent Means of Positive and Negative Impression Management

Positive IM in scalar invariance model		**Positive IM in partial scalar invariance model**		**Negative IM in scalar invariance model**		**Negative IM in partial scalar invariance model**	
*Country**	*Latent Mean*	*Country**	*Latent Mean*	*Country**	*Latent Mean*	*Country**	*Latent Mean*
Abu Dhabi (UAE)	.38	Abu Dhabi (UAE)	.39	Korea	.70	Korea	.76
Serbia	.15	Serbia	.16	Latvia	.68	Latvia	.42
Mexico	.12	Mexico	.12	Poland	.57	Estonia	.34
Portugal	.11	Portugal	.12	Estonia	.49	Poland	.31
Brazil	.00	Brazil	.00	Finland	.45	Chile	.30
Spain	-.02	Spain	-.03	France	.36	Finland	.28
Iceland	-.04	Iceland	-.07	Chile	.35	Slovak Rep.	.27
Croatia	-.07	Croatia	-.07	Malaysia	.32	France	.10
Chile	-.13	Chile	-.13	United States	.27	United States	.09
United States	-.15	United States	-.15	Slovak Rep.	.26	Malaysia	.09
Latvia	-.27	Latvia	-.27	Iceland	.18	Croatia	.08
Poland	-.29	Poland	-.28	Spain	.08	Iceland	.07
Slovak Rep.	-.29	Slovak Rep.	-.29	Croatia	.08	Brazil	.00
France	-.30	Estonia	-.31	Portugal	.07	Spain	-.05
Estonia	-.31	Malaysia	-.34	Mexico	.01	Mexico	-.09
Malaysia	-.38	France	-.34	Brazil	.00	Portugal	-.13
Finland	-.46	Finland	-.47	Serbia	-.03	Serbia	-.25
Korea	-.87	Korea	-.86	Abu Dhabi (UAE)	-.12	Abu Dhabi (UAE)	-.26

Note: * Countries were ranked based on their scores on the latent mean on each dimension.

The correlation of the country scores of positive impression management from the two sets of estimations was 1, and that of negative impression management was .91. The rank orders of countries based on scores of positive impression management in the two sets were mostly the same. The rank orders based on the two sets of negative impression management scores were less stable, especially the changes in positions of countries ranked in the middle were rather large.

In the *observed mean* approach, the scale scores of positive and negative impression management were computed, each with five items or with four items (as suggested in the partial invariance model that Item 9 "I admit when I do not know something if a student asks a question in class" was dropped from positive impression management, and Item 10 "I am irritated by students who ask for favours" from the negative impression management). Internal consistency coefficients (Cronbach's alpha) of the scales were calculated. In the first case, the reliability was marginally acceptable, with values of .59 for the 5-item positive impression management scale, .48 for the 5-item negative impression management scale, and .61 if the two subscales were taken as one scale (with the items of the negative impression management scale reverse keyed). In the latter case (using four items), the scales showed improved reliability, with values of .66, .52, and .65, respectively. Table 6 shows the reliability values for each country in both solutions.

Table 6 Values of Cronbach's Alpha for Positive and Negative Impression Management in the 5-Item and 4-Item Solutions

Country	**Positive IM** *(5 items)*	**Negative IM** *(5 items)*	**Global** *(10 items)*	**Positive IM** *(4 items)*	**Negative IM** *(4 items)*	**Global** *(8 items)*
Abu Dhabi (UAE)	.50	.53	.58	.61	.52	.60
Brazil	.59	.45	.56	.60	.39	.52
Chile	.68	.75	.76	.69	.72	.73
Croatia	.57	.53	.63	.63	.54	.65
Estonia	.61	.54	.65	.63	.54	.65
Finland	.66	.54	.68	.64	.51	.67
France	.47	.46	.54	.49	.50	.57
Iceland	.63	.49	.63	.62	.45	.60
Korea	.75	.70	.76	.73	.64	.70
Latvia	.59	.53	.61	.66	.56	.65
Malaysia	.45	.54	.53	.68	.53	.63
Mexico	.57	.52	.59	.60	.51	.59
Poland	.67	.49	.62	.69	.49	.64
Portugal	.47	.39	.47	.61	.50	.60
Serbia	.59	.40	.54	.68	.51	.65
Slovak Rep.	.60	.59	.67	.57	.56	.62
Spain	.61	.51	.64	.60	.49	.62
United States	.67	.53	.68	.67	.51	.67
Overall	*.59*	*.48*	*.61*	*.65*	*.52*	*.65*

Note: IM = Impression Management

The correlation between the scores of the 5-item and 4-item positive impression management at the individual level (i.e., respondent-by-respondent basis) was .91, and that of negative impression management was .90. At the country level (i.e., with the individual-level data aggregated at the country level), the correlations were .80 and .75, respectively. A Multivariate Analysis of Variance (MANOVA) was performed with the scale scores of positive and negative impression management as dependent variables, and country as the independent variable. The effect sizes were compared between the 5-item and 4-item solutions. The eta square of positive impression management changed from .11 to .09, and that of negative impression management changed from .11 to .09, which indicated inconsequential differences in the size of cross-cultural differences. The rank orders of countries were slightly different in the two solutions. The rank orders and the scale scores are presented in Table 7. In the case of positive impression management, higher scores (scores closer to 7) indicate a stronger tendency to attribute positive traits and behaviors to oneself, whereas in the case of negative impression management lower scores (scores closer to 1) indicate a stronger tendency to deny negative traits or behaviours. So, higher scores on the positive impression management and lower scores on the negative impression management scales are indicative of more impression management. Similar to the patterning in country ranking according to the latent means, in the observed mean approach, teachers in Korea had a weak tendency to engage in positive impression management and a strong tendency in negative impression management.

Table 7 Country Rankings and Observed Means of Positive and Negative Impression Management

Positive IM (5 items)		**Positive IM (4 items)**		**Negative IM (5 items)**		**Negative IM (4 items)**	
*Country**	*Scale Mean*	*Country**	*Scale Mean*	*Country**	*Scale Mean*	*Country**	*Scale Mean*
Abu Dhabi (UAE)	6.34	Abu Dhabi (UAE)	6.55	Abu Dhabi (UAE)	1.69	Serbia	1.63
Mexico	6.24	Serbia	6.29	Mexico	1.87	Abu Dhabi (UAE)	1.63
Portugal	6.20	Portugal	6.28	Croatia	1.97	Croatia	1.79
Serbia	6.19	Mexico	6.26	Spain	2.01	Portugal	1.83
Brazil	6.19	Brazil	6.18	Brazil	2.08	Mexico	1.86
Spain	6.18	Spain	6.13	Iceland	2.11	Spain	1.97
Iceland	6.18	Croatia	6.10	Slovak Rep.	2.18	Iceland	2.08
Croatia	6.09	Iceland	6.07	United States	2.25	United States	2.12
United States	6.07	Chile	6.03	Serbia	2.25	France	2.13
Chile	6.01	United States	6.01	Chile	2.28	Malaysia	2.19
Poland	5.91	Latvia	5.92	Malaysia	2.39	Brazil	2.19
Latvia	5.91	Malaysia	5.89	Finland	2.49	Slovak Rep.	2.27
Estonia	5.89	Poland	5.87	Estonia	2.52	Chile	2.31
Slovak Rep.	5.88	Estonia	5.85	Portugal	2.57	Finland	2.36
France	5.88	Slovak Rep.	5.82	France	2.57	Estonia	2.42
Finland	5.84	France	5.74	Poland	2.76	Poland	2.43
Korea	5.40	Finland	5.72	Latvia	2.84	Latvia	2.43
Malaysia	5.25	Korea	5.34	Korea	2.97	Korea	3.08

Note. IM = impression management.

* Countries were ranked based on their scores on the observed mean on each dimension.

So, the teacher social desirability data from these 18 countries demonstrated acceptable metric invariance and excellent partial metric invariance (two items showed poor invariance in loadings and intercepts: Item 9 "I admit when I do not know something if a student asks a question in class" and Item 10 "I am irritated by students who ask for favours") (See Table 4). Scalar invariance was not supported by the fit statistics (See Table 4). However, the latent country means as well as observed scale scores based on scalar and partial scalar models do not seem to show large differences and the size of cross-country differences were similar with the two solutions (See Table 5 and 7). Given the significantly better fit of the partial invariance model, and the higher reliability of scales with only the invariant items (See Table 6), we decided to base the remaining analyses on the partial invariance model (i.e., using the 4-item positive and the 4-item negative impression management scale scores).

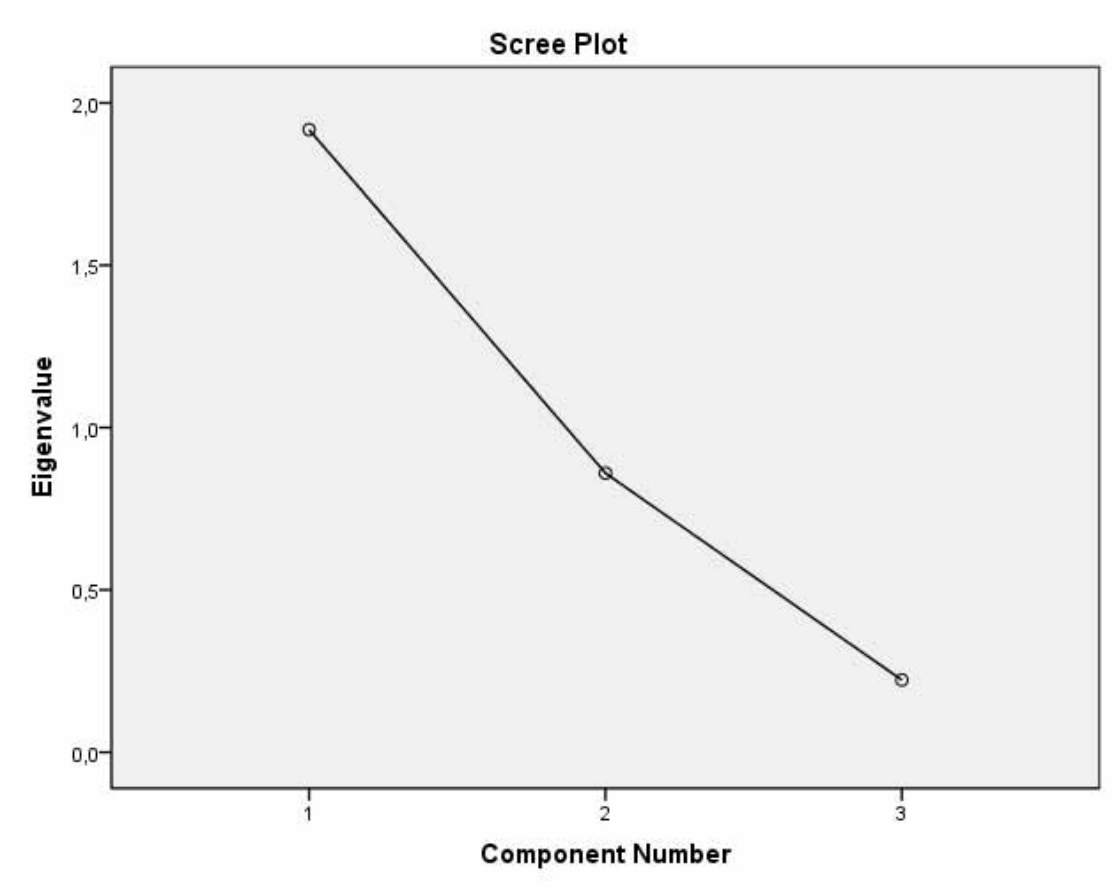
Box 1: Synopsis of Measurement Invariance Testing

- This study supports the view that social desirability scale has two sub dimensions: positive and negative impression management.
- The internal consistencies of the positive and negative impression management scales are modest.
- Across the 18 countries, the two factors show excellent partial metric invariance, suggesting that the same constructs are measured and the items contribute similarly to these constructs.
- The partial scalar invariance model fits significantly better than the full scalar invariance model, yet the full and partial scalar invariance model produce very similar patterns of cross-cultural differences. With 18 countries in total, conclusions about cross-cultural differences in mean scores do not seem to be affected by the problematic invariance of the scores across the countries.
- Given the significantly better fit of the partial invariance model, and the higher reliability of scales with only the invariant items, scale scores of positive and negative impression management based on the partial invariance model are used.

3.2 Integrating Specific Response Styles in a General Response Style Factor

A general response style factor (GRS), explaining 64% of all the variance, was extracted in a principal component analysis of the 8-item social desirability scale (the negative impression management items reversed and combined with the positive impression management scores), extreme and midpoint responding. The scree plot (depicted in Figure 2) supported a one-factor solution. As expected, social desirability (.51) and extreme responding (.91) loaded positively and midpoint responding (-.91) loaded negatively on the factor. The presence of a general response style suggests that the three response styles studied here—social desirability, extreme and midpoint responding—can be viewed as indicators of a common underlying tendency. In a previous study we interpreted the general response style as a communication filter, which represents the tendency of response amplification (e.g., using more social desirability and extreme responding) versus response moderation (e.g., using more midpoint responding) (He & van de Vijver, 2013). It should be noted that the general response style factor captures the shared meaning in the specific response styles, whereas each response style still has some uniqueness. For example, extremity scoring may not be the exact opposite of midpoint responding, but the strong negative, yet imperfect correlations between the two tendencies point to shared meaning as well as some uniqueness of each of the two response styles.

Figure 2 Scree Plot in the Exploratory Factor Analysis of Social Desirability, Extreme and Midpoint Responding



Source: Author's own work using TALIS 2013 data.

Box 2: Synopsis of Integration of Response Styles

- Social desirability, extremity and midpoint responding can be integrated in a general response style factor.
- Social desirability and extreme responding load positively and midpoint responding loads negatively on the general response style factor. This general factor presumably represents the tendency to either moderate or amplify responses. This general response style factor comprises the shared meaning in the three specific response styles, which is an important tool in furthering our understanding of response styles.

3.3 Country-Level Correlates of Social Desirability and the General Response Style

Country scores on positive and negative impression management and the general response style were correlated with a host of socioeconomic and education indicators (all indicators were available for all countries unless indicated otherwise):

- The *Better Life Index* measures well-being across countries based on 11 topics that OECD has identified as essential in the areas of material living conditions and quality of life. It was calculated as the mean of the standardised scores of 24 indicators. The data were taken from the 2011 OECD report (http://stats.oecd.org/Index.aspx?DataSetCode=BLI). Data on this index are available for 13 countries in the present study, but not available for Croatia, Malaysia, Latvia, Serbia, and Abu Dhabi (UAE).

- The *Human Development Index* is a composite measure of the average achievements in three basic dimensions of human development: a long and healthy life, access to knowledge (schooling), and a high standard of living. Data of the Human Development Index of 2012 were taken from the UNDP (https://data.undp.org/dataset/Table-1-Human-Development-Index-and-its-components/wxub-qc5k).

- Students' performance in reading, mathematics and science indicated as mean scores of reading, mathematics, and science in PISA 2012 were taken from the OECD (www.oecd.org/pisa/keyfindings/PISA-2012-results-snapshot-Volume-I-ENG.pdf).

- Scores of satisfaction with education (i.e., percentage of respondents who answered "satisfied" to the Gallup World Poll question, "Are you satisfied or dissatisfied with the education system? ") were taken from the United Nations Development Programme (https://data.undp.org/dataset/Table-8-Education/mvtz-nsye). Data were available for all countries except Iceland.

Table 8 below shows the correlations of positive and negative impression management and the general response style with these country-level indicators. Positive impression management and general response style was negatively correlated with HDI and achievement in reading, mathematics and science, whereas negative impression management showed the opposite patterning. It suggests that respondents in countries with higher levels of socioeconomic development and educational achievement are on average less inclined to show stronger levels of impression management than respondents in countries with lower levels of socioeconomic development and educational achievement.

Table 8 Correlation of Response Styles and Social Indicators at Country Level

Country-level Indicator of Socioeconomic Development and Education Achievement	***PIM***	***NIM***	***GRS***
Better Life Index (n = 13 countries)	-.27	.08	.06
Human Development Index (n = 18)	-.49*	.36	-.23
PISA Reading Mean Score (n = 18)	-.59*	.48*	-.56*
PISA Mathematics Mean Score (n = 18)	-.67**	.56*	-.69**
PISA Science Mean Score (n = 18)	-.61**	.52*	-.63**
Satisfaction with Education (n = 17)	.21	-.35	.24

Note : PIM = positive impression management. NIM = negative impression management; GRS = general response style. *p < .05; **p < .01.

Since the Better Life Index is a composite variable from 25 dimensions covering 11 topics (the dimensions can be found on http://stats.oecd.org/Index.aspx?DataSetCode=BLI), a further analysis was performed to study which dimensions in the Better Life Index would explain response styles more. A step-wise regression was carried out with positive impression management as the dependent variable and the 25 individual dimension scores as the independent variables. *Educational Attainment* (standardised β = -.71) and *Self-Reported Health* (β = .43) could explain in total 69% of the variance. The same analysis was done for negative impression management, which was predicted by *Educational Attainment* (β = .68), *Self-*

Reported Health (β = -.87), and *Life Satisfaction* (β = .45), explaining 82% of the variance. For the general response style factor, the best predictors were *Educational Attainment* (standardised β = -.67) and *Self-Reported Health* (β = .54), explaining 73% of the variance. These findings suggest that some aspects of the Better Life Index are more predictive of social desirability than other aspects. Social desirability is also known to be related to education at individual level.

Box 3: Synopsis of Correlations with Country Characteristics

- Positive impression management at the country level is negatively correlated with indicators of socioeconomic development and educational achievement of countries, suggesting that countries with less affluence and lower education development tend to have higher scores on positive impression management.
- The opposite pattern is observed for negative impression management: a positive correlation was found, suggesting that countries with higher affluence and higher educational achievement tend to have higher scores on negative impression management (note that in this study higher scores on this factor are indicative of less impression management).
- The general response style, like positive impression management, is negatively correlated with educational achievement at country level.
- Findings suggest that in countries with higher levels of economic development and educational achievement respondents are less inclined to demonstrate the studied response styles than respondents in countries with lower levels of economic development and educational achievement.

3.4 Implications of Corrections for Social Desirability

We report the correlations of the two social desirability factors with core constructs in TALIS at both individual and country level, in order to check which constructs are more susceptible to the influence of response styles. Then, we (1) compare the effect sizes of cross-cultural differences before and after controlling for these social desirability factors; and (2) calculate the correlations of the core constructs before and after controlling for these social desirability factors to examine to what degree social desirability may affect the cross-cultural differences.

An overview of the 14 individual constructs (Three composite constructs are constructed from some of these individual constructs) is presented in Table 9. All scales showed acceptable reliability; measurement invariance of each scale across countries was tested in multigroup confirmatory factor analyses. As documented in the technical report (OECD, 2014), configural invariance (i.e., the same indicators loaded on the same latent variables across countries) was supported in all scales; metric invariance (i.e., factor loadings constrained to be equal across countries) was supported in most scales except a marginally acceptable fit on Satisfaction with current work environment, and scalar invariance (i.e., items constrained to have the same intercepts across countries) was rarely supported. However, in many cases, factor scores of the scale computed from scalar invariance showed very strong correlations with those based on metric invariance model, pointing to similar robustness of the cross-cultural differences between the two invariance models. For scales showing similarity in the two invariance models, factor scores of the scales were computed from the scalar invariance model; otherwise these were computed from the metric invariance model.

Table 9 Overview of the Core Constructs in TALIS

Scale	Subscales(if any)	No of items	Cronbach's Alpha	Invariance level used for computing factor scores
Teacher efficacy	Efficacy in classroom management	4	.82	Scalar
	Efficacy in instruction	4	.75	Scalar
	Efficacy in student engagement	4	.78	Scalar
Job satisfaction	Satisfaction with current work environment	4	.78	Metric
	Satisfaction with teaching profession	4	.81	Metric
Teacher cooperation	Exchange and coordination for teaching	4	.72	Scalar
	Professional collaboration	4	.60	Scalar
Participation among stakeholders		5	.85	Scalar
Teacher-student relations		4	.78	Scalar
Classroom disciplinary climate		4	.85	Metric
Constructivist beliefs		4	.71	Metric
Effectiveness in professional development		4	.70	Scalar
Professional development in subject matter/ pedagogy		5	.84	Metric
Professional development for diversity		6	.82	Metric

3.4.1 Correlations of social desirability and core constructs

The correlations between positive and negative impression management and core constructs at the individual level (i.e., respondent-by-respondent basis) and at the country level (i.e., with the individual level data aggregated at the country level) are presented in Table 10. These core constructs (shown in Table 10) were assessed with 17 scales (the scale properties and invariance testing were documented in the forthcoming TALIS technical report). Constructs related to teacher's efficacy including efficacy in classroom management, instruction, and student engagement showed the strongest correlations with positive and negative impression management at both the individual and country level, followed by constructs related to different aspect of job satisfaction, which suggests that not all TALIS constructs assessed by the questionnaire show the same susceptibility to social desirability. The correlations at the country level were stronger than these at the individual level, presumably because random errors were smaller after data were aggregated from the individual level to the country level.

Box 4: Synopsis of Correlations of Social Desirability Scores with Core TALIS Constructs

- At the individual level, positive impression management is positively correlated with aspects of teacher's efficacy and job satisfaction; not surprisingly, negative impression management shows the opposite patterning.
- Country-level correlations of response styles with core constructs are similar, yet stronger.
- Constructs like teacher efficacy and job satisfaction assessed by the questionnaire are most susceptible to social desirability.

Table 10 Correlations of Positive and Negative Impression Management with the Core Constructs in TALIS at the Individual and Country Level

Scale	Individual-level Positive IM	Individual-level Negative IM	Country-level Positive IM	Country-level Negative IM
Efficacy in classroom management	.30**	-.20**	.50*	-.50*
Efficacy in instruction	.39**	-.26**	.77**	-.72**
Efficacy in student engagement	.34**	-.19**	.50*	-.32
Overall teacher efficacy	.37**	-.23**	.65**	-.56*
Satisfaction with current work environment	.24**	-.21**	.56	-.54
Satisfaction with teaching profession	.18**	-.21**	.20	-.35
Overall job satisfaction	.23**	-.23**	.39	-.47*
Participation among stakeholders	.13**	-.07**	.03	.03
Teacher-student relations	.21**	-.15**	.34	-.32
Classroom disciplinary climate	.10**	-.14**	.15	-.25
Constructivist beliefs	.10**	-.05**	.24	-.14
Exchange and coordination for teaching	.15**	-.10**	.32	-.34
Professional collaboration	.14**	-.04**	.43	-.19
Teacher cooperation	.16**	-.08**	.53*	-.36
Effectiveness in professional development	.12**	-.06**	.56*	-.39
Professional development in subject matter/ pedagogy	-.15**	.19**	-.47*	.41
Professional development for diversity	-.02**	.10**	-.08	.15

Note: IM = impression management.

* $p < .05$; ** $p < .01$.

3.4.2 Correcting for social desirability

The influence of social desirability on key concepts in the TALIS teacher questionnaire was assessed in a series of Analyses of Covariance (ANCOVAs). Each of the 17 core constructs scales measured in TALIS served as a dependent variable, country was the group variable, and the positive and/or negative impression management at individual level was/were entered as covariates. The effect sizes of the cross-cultural differences on these scales before and after correcting for positive impression management, negative impression management, and both combined were compared. Technically, this procedure is identical to a regression procedure in which the core TALIS constructs are used as dependent variables and social desirability is used as an independent variable, and residuals (i.e., scores on the core constructs after correction for social desirability) are then tested for country differences.

One of the most important assumptions for an ANCOVA is the homogeneity of regression coefficients of the covariates. These are the coefficients of the line, regressing the TALIS construct scale score on social desirability. If the regression coefficients (slopes) of the covariate are different across groups, the correction of the covariate could result in incorrect results. We tested the homogeneity of regression coefficients of the covariates across countries on the dependent variables in multilevel analyses before carrying out ANCOVAs. We tested the significance and the effect sizes of the interaction of individual- and country-level positive and negative impression management scores (standardised) on the standardised scores of the 17 core TALIS constructs in a random slope and random intercepts model. If the interaction of individual- and country-level covariates is non-significant or has very small effect sizes, it indicates that the homogeneity of regression coefficients is not violated and an ANCOVA will produce accurate estimates.

Seven out of 17 cross-level interactions of positive impression management were non-significant; for the significant interactions, the effect sizes were very small (standardised regression ranged from -.033 to .026). Eight of the 17 cross-level interactions of negative impression management were non-significant, and the effect sizes for the significant interactions also were small (standardised regression ranged from -.030 to .014). We conclude that the homogeneity of regression coefficients of the covariates was fairly well supported; therefore, we decided that ANCOVAs would be appropriate.

In Table 11 below effect sizes refer to proportions of variance accounted for by country in analyses of (co)variance before and after correction. The last two columns give the correlations between the corrected and uncorrected scores at the individual and country level. Correlation values closer to 1 indicate that the correction has less influence.

Table 11 The Effects of Correcting for Social Desirability in the Core TALIS Constructs

Scale	Effect size before correction	Effect size after Positive IM correction	Effect size after Negative IM correction	Effect size after both Positive and Negative IM correction	Correlation individual scores Positive and Negative IM combined corrected-uncorrected	Correlation country scores Positive and Negative IM corrected-uncorrected
Efficacy in classroom management	.11	.10	.10	.10	.95	.95
Efficacy in instruction	.21	.15	.17	.15	.92	.96
Efficacy in student engagement	.27	.25	.25	.25	.95	.97
Overall teacher efficacy	.18	.15	.16	.15	.93	.95
Satisfaction with current work environment	.06	.05	.05	.04	.96	.92
Satisfaction with teaching profession	.14	.15	.14	.14	.97	.96
Overall job satisfaction	.10	.09	.09	.09	.96	.93
Participation among stakeholders	.04	.04	.04	.05	.99	.98
Teacher-student relations	.05	.05	.05	.05	.98	.95
Classroom disciplinary climate	.10	.11	.11	.11	.99	.98
Constructivist beliefs	.07	.06	.06	.06	1.00	.99
Exchange and coordination for teaching	.24	.24	.24	.24	.99	.99

Scale	Effect size before correction	Effect size after Positive IM correction	Effect size after Negative IM correction	Effect size after both Positive and Negative IM correction	Correlation individual scores Positive and Negative IM combined corrected-uncorrected	Correlation country scores Positive and Negative IM corrected-uncorrected
Professional collaboration	.30	.27	.27	.27	.99	1.00
Teacher cooperation	.13	.12	.12	.12	.99	.99
Effectiveness in professional development	.05	.05	.05	.05	.99	.99
Professional development in subject matter/ pedagogy	.21	.19	.19	.19	.98	.99
Professional development for diversity	.23	.24	.24	.24	1.00	1.00

Note. All effect sizes and correlations are significant ($p < .05$). IM = impression management.

The average country effect size before correction is large (about .144) (Cohen, 1988) and reduces just slightly (to about .136) after correction for both positive and negative impression management. It indicates that the two scales of social desirability do not reduce the effect sizes much. Correlations between scores before and after correction for social desirability are .97 both at individual and country level. Therefore, correction for social desirability does not have a sizeable impact on the size of the country differences observed.

Table 12 below summarises countries with the highest and lowest scores on each of the target scales before and after correcting for both positive and negative impression management. As can be seen, except in the cases of five constructs--Participation among stakeholders, Teacher cooperation, Professional collaboration, Professional development in subject matter/ pedagogy, and Professional development for diversity --wherein the rank orders of the top and bottom three countries remained the same before and after correction, the rank orders of countries on all 12 other TALIS scale constructs changed slightly after correction, which points to a modest impact of social desirability on different target scales.

Table 12 Top Three and Bottom Three Countries Before and After Correction of Social Desirability in Each TALIS Core Construct

Scale	Top three before correction	Top three after correction	Bottom three before correction	Bottom three after correction
Efficacy in classroom management	Abu Dhabi (UAE) Malaysia Portugal	Malaysia Abu Dhabi (UAE) France	Spain Estonia Korea	Estonia Korea Spain
Efficacy in instruction	Abu Dhabi (UAE) Portugal Brazil	Portugal Abu Dhabi (UAE) Malaysia	Finland Estonia Korea	Finland Korea Estonia
Efficacy in student engagement	Abu Dhabi (UAE) Portugal Malaysia	Abu Dhabi (UAE) Malaysia Portugal	Spain Korea Croatia	Serbia Spain Croatia
Overall teacher efficacy	Abu Dhabi (UAE) Portugal Malaysia	Abu Dhabi (UAE) Portugal Malaysia	Croatia Estonia Korea	Korea Estonia Croatia
Satisfaction with current work environment	Mexico Iceland Portugal	Mexico Iceland Chile	Slovak Rep. Estonia Korea	Estonia Abu Dhabi (UAE) Korea
Satisfaction with teaching profession	Mexico Malaysia Spain	Mexico Malaysia Finland	Korea Latvia Slovak Rep.	Portugal Brazil Slovak Rep.
Overall job satisfaction	Mexico Malaysia Spain	Mexico Malaysia Finland	Estonia Slovak Rep. Korea	Estonia Abu Dhabi (UAE) Slovak Rep.
Participation among stakeholders	Latvia Malaysia Poland	Latvia Malaysia Poland	France United States Mexico	France United States Mexico
Teacher-student relations	Iceland Abu Dhabi (UAE) United States	Iceland United States Chile	Brazil Slovak Rep. Korea	Slovak Rep. Mexico Brazil

Scale	Top three before correction	Top three after correction	Bottom three before correction	Bottom three after correction
Classroom disciplinary climate	Abu Dhabi (UAE)	Poland	Spain	Spain
	Croatia	Estonia	Chile	Chile
	Serbia	Croatia	Brazil	Brazil
Constructivist beliefs	Abu Dhabi (UAE)	Korea	Malaysia	Estonia
	Korea	Abu Dhabi (UAE)	Estonia	Brazil
	Mexico	Mexico	Spain	Spain
Exchange and coordination for teaching	Spain	Spain	Brazil	Brazil
	Portugal	Poland	Mexico	Mexico
	Poland	Portugal	Korea	Korea
Professional collaboration	Abu Dhabi (UAE)	Abu Dhabi (UAE)	Finland	Finland
	Mexico	Mexico	Spain	Spain
	Poland	Poland	France	France
Teacher cooperation	Abu Dhabi (UAE)	Abu Dhabi (UAE)	Croatia	Croatia
	Poland	Poland	France	France
	Estonia	Estonia	Korea	Korea
Effectiveness in professional development	Abu Dhabi (UAE)	Abu Dhabi (UAE)	Iceland	Iceland
	United States	United States	Slovak Rep.	Korea
	Malaysia	Latvia	Korea	Slovak Rep.
Professional development in subject matter/ pedagogy	Malaysia	Malaysia	United States	United States
	Korea	Korea	Abu Dhabi (UAE)	Abu Dhabi (UAE)
	Iceland	Iceland	Poland	Poland
Professional development for diversity	Korea	Korea	United States	United States
	Brazil	Brazil	Finland	Finland
	Malaysia	Malaysia	Poland	Poland

We conclude with a caveat. Score corrections for the core TALIS constructs need to take into consideration if these constructs show scalar invariance themselves. So, in order to study the effect of score corrections for social desirability, we need to assume that the core TALIS constructs are not challenged by invariance issues and show scalar invariance (the measurement invariance of these core constructs are documented in the TALIS technical report) (OECD, 2014). Scalar invariance of the TALIS scales should be seen here as a necessary assumption. According to the TALIS 2013 Technical Report (OECD, 2014), there were various scales that showed metric invariance, but no scale showed scalar invariance. Some

scales came rather close (such as classroom disciplinary climate), whereas teacher cooperation did not meet criteria for scalar invariance in any way. As demonstrated in Table 9, in many cases the country means computed for partial and full scalar invariance of the TALIS scales show the same cross-cultural patterning. Thus, our analyses should be accurate.

Box 5: Synopsis of Implications of Correction for Social Desirability

- Corrections for positive and negative impression management do not measurably change the effect sizes of cross-cultural differences in the 17 core TALIS constructs, including aspects of teacher's efficacy, job satisfaction, and school climate.
- The rank orders of countries on core TALIS constructs before and after correcting for the two social desirability factors are generally stable. This means that social desirability does not affect the validity of country differences in these TALIS constructs.

3.5 Implications of Corrections for the General Response Style

3.5.1 Correlations of the general response style with core TALIS constructs

The analysis of the previous section was extended to the generalised response tendency by correlating the general response style with the 17 core TALIS constructs at both individual and country level (Table 13). A similar correlation patterning was found as in the case of positive impression management, with differential correlations with the core constructs and the strongest associations with teacher's efficacy and job satisfaction constructs at both levels.

Table 13 Correlations of the General Response Style with the Core Constructs in TALIS at the Individual and Country Level

Scale	GRS individual level	GRS country level
Efficacy in classroom management	.33**	.49*
Efficacy in instruction	.41**	.73**
Efficacy in student engagement	.35**	.47*
Overall teacher efficacy	.39**	.61**
Satisfaction with current work environment	.41**	.77**
Satisfaction with teaching profession	.36**	.56*
Overall job satisfaction	.42**	.72**
Participation among stakeholders	.15**	-.34
Teacher-student relations	.42**	.50*
Classroom disciplinary climate	.18**	-.01
Constructivist beliefs	.29**	.29
Exchange and coordination for teaching	.13**	.15
Professional collaboration	.10**	.17
Teacher cooperation	.13**	.22
Professional development in subject matter/ pedagogy	.14**	.44
Effectiveness in professional development	-.14**	-.33
Professional development for diversity	-.02**	.04

Note: GRS = General Response Style. *$p < .05$; **$p < .01$.

3.5.2 Correcting for the general response style

Scores on the general response style factor were entered as a covariate in a series of ANCOVAs, with country as the independent variable and the core TALIS constructs as dependent variables. The homogeneity of the regression coefficients of the general response style factor on the 17 core constructs was checked; the standardised regression coefficients of the cross-level interaction ranged from -.047 to .010, which indicated very small effect sizes. The effect sizes and correlations before and after correcting for the general response style factor are presented in Table 14 below. A similar conclusion to the analyses with positive and negative impression management can be drawn based on the similarity of effect sizes and high correlations before and after the correction: the response style correction does not seem to affect the observed cross-cultural differences and the rank order of countries on the core TALIS constructs.

Table 14 Effects of Correcting for the General Response Style in the Core TALIS Constructs

Scale	Effect size before correction	Effect size after GRS correction	Correlation individual scores corrected-uncorrected GRS	Correlation country scores corrected-uncorrected GRS
Efficacy in classroom management	.11	.11	.95	.96
Efficacy in instruction	.21	.17	.91	.97
Efficacy in student engagement	.27	.26	.93	.98
Overall teacher efficacy	.18	.17	.92	.96
Satisfaction with current work environment	.06	.03	.91	.90
Satisfaction with teaching profession	.14	.13	.93	.95
Overall job satisfaction	.10	.07	.91	.92
Participation among stakeholders	.04	.05	.99	.98
Teacher-student relations	.05	.06	.91	.88
Classroom disciplinary climate	.10	.11	.98	.98
Constructivist beliefs	.07	.06	.96	.93
Exchange and coordination for teaching	.24	.24	.99	1.00
Professional collaboration	.30	.27	1.00	1.00
Teacher cooperation	.13	.12	.99	.99
Effectiveness in professional development	.05	.05	.99	.99
Professional development in subject matter/ pedagogy	.21	.19	.99	1.00
Professional development for diversity	.23	.24	1.00	1.00

Note: GRS = General Response Style

Box 6: Synopsis of Implications of Correction for General Response Style

- The general response style has differential correlations with the 17 core TALIS constructs. It was most strongly associated with teacher's efficacy and job satisfaction constructs, and least associated with constructs related to professional development and teacher cooperation.
- Controlling for the general response style does not measurably change the effect sizes of cross-cultural differences in the 17 core constructs.
- Controlling for the general response style does not meaningfully change the rank order of countries on this core constructs.

PART 4: CONCLUSIONS

The study of response styles, notably social desirability, has been controversial in psychology. On the one hand, the idea has been advocated that such response styles present distorted representations of participants' views. A good example of this approach can be found in the work by Eysenck and Eysenck (1975), who proposed to interpret scores on personality scales only if a participant's score on a social desirability scale was below a pre-determined critical threshold. On the other hand, the idea has been proposed that social desirability is part of agreeableness (McCrae & Costa, 1983). There has also been some controversy on the need to correct for response styles in survey research. Several authors have argued that corrections for response styles do not have a serious impact on applications of test scores. For example, the validity of an instrument to predict job performance is not strongly influenced by a correction for response styles (Ones, Viswesvaran, & Reiss, 1996).

The present study addressed the meaning and implications of response styles, notably social desirability, for the TALIS 2013 data. With 18 of the 34 countries that participated in TALIS, there was fairly good cultural variation in our sample. For example, data from both Central / South America and East Asian countries were available; these countries are known to differ considerably on midpoint and extreme responding.

The analysis of invariance of the scale showed strong support for the identity of the underlying social desirability factors, positive and negative impression management. Although scalar invariance was not supported, we found that comparisons of all items and items that are least affected by bias yielded a similar pattern of cross-cultural differences. These findings suggest that despite the absence of a solid statistical basis for country-level comparisons of scores, such comparisons are unlikely to yield a highly distorted picture.

The TALIS constructs that were most related to social desirability (and the combination of social desirability, midpoint responding, and extreme responding in a general response style) were teacher efficacy and teacher job satisfaction. These findings suggest that response styles do not affect TALIS constructs in the same manner. There is evidence that response styles are triggered most in questions about personal domains when evaluation apprehension is strongest (van Dijk, Datema, Piggen, Welten, & van de Vijver, 2009). Teachers may feel more personal involvement in evaluating their efficacy and job satisfaction than their professional development. In addition to domain differences, we also found country differences in response styles. We found that, notably socioeconomic and educational indicators, such as the Human Development Index, and math, science, and reading achievement scores showed strong correlations with the two factors of social desirability. In general, it seems that countries with high levels of economic development and high scores on educational achievement showed low scores on social desirability (and the general response style).

We used analysis of covariance-based correction procedures for studying the impact of response styles on the size of the observed cross-cultural differences. The impact of score corrections was remarkably small; this was found for both constructs that showed very weak and very strong correlations with the response style. This result seems counterintuitive, even if as argued above, it has been observed before (e.g., Ones et al., 1996). The similarity of effect sizes of country differences before and after correction for response styles means that the correction leaves the country differences intact. The score

rank order and the relative differences of country scores are not affected by these corrections. The same was observed at the individual level. It is unlikely that score corrections for response styles will have a major impact on individual or country differences of core TALIS constructs. It seems fair to conclude that response styles, including social desirability, midpoint responding, and extreme responding, have a detectable bearing on responses on the TALIS survey, notably at country level. However, "peeling off" the influence of social desirability (or the general response styles for that matter) does not change the country order of means substantially. For future TALIS rounds, our findings suggest that it may be sufficient to use items measuring different constructs in TALIS to construct indirect measures of response styles (e.g., extreme and midpoint responding), which have a similar working mechanism as social desirability. In addition, multiple methodologies such as controlling for the effects of respondents' over claiming, rescaling based on anchoring vignettes, and decomposing country differences by language groups could be pursued to shed additional light on potential bias on cross-cultural surveys.

REFERENCES

Austin, E. J., Deary, I. J., & Egan, V. (2006). Individual differences in response scale use: Mixed Rasch modelling of responses to NEO-FFI items. Personality and Individual Differences, 40, 1235-1245. doi:10.1016/j.paid.2005.10.018

Barger, S. D. (2002). The Marlowe-Crowne affair: Short forms, psychometric structure, and social desirability. Journal of Personality Assessment, 79, 286-305. doi:10.1207/s15327752jpa7902_11

Baumgartner, H., & Steenkamp, J.-B. E. M. (2006). Response biases in marketing research. In R. Grover & M. Vriens (Eds.), The handbook of marketing research: : Uses, misuses, and future advances (pp. 95-109). Thousand Oaks, CA: Sage.

Byrne, B. M., & van de Vijver, F. J. R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. International Journal of Testing, 10, 107-132. doi:10.1080/15305051003637306

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. Journal of Consulting Psychology, 24, 349-354.

Crowne, D. P., & Marlowe, D. (1964). The approval motive: studies in evaluative dependence. New York, NY: John Wiley and Sons.

De Beuckelaer, A., Weijters, B., & Rutten, A. (2010). Using ad hoc measures for response styles: A cautionary note. Quality & Quantity, 44, 761-775. doi:10.1007/s11135-009-9225-z

de Vries, R. E., Zettler, I., & Hilbig, B. E. (2013). Rethinking trait conceptions of social desirability scales: Impression management as an expression of honesty-humility. Assessment. doi:10.1177/1073191113504619

Eysenck, H. J., & Eysenck, S. B. G. (1975). Manual of the Eysenck Personality Questionnaire. London: Hodder and Stoughton.

He, J., van de Vijver, F., J. R., Domínguez, A. d. C., & Mui, P. H. C. (in press). Acquiescent, extreme, and midpoint response styles: A multilevel study. International Journal of Cross-Cultural Management.

He, J., & van de Vijver, F. J. R. (2013). A general response style factor: Evidence from a multi-ethnic study in the Netherlands. Personality and Individual Differences, 55, 794-800. doi:10.1016/j.paid.2013.06.017

Johnson, T. P., & van de Vijver, F., J. R. (2003). Social desirability in cross-cultural research. In J. A. Harkness, F. J. R. van de Vijver & P. P. Mohler (Eds.), Cross-cultural survey methods (pp. 195-204). Hoboken, New Jersey: John Wiley & Sons.

Loo, R., & Loewen, P. (2004). Confirmatory factor analyses of scores from full and short versions of the Marlowe–Crowne Social Desirability Scale. Journal of Applied Social Psychology, 34, 2343-2352. doi:10.1111/j.1559-1816.2004.tb01980.x

McCrae, R. R., & Costa, P. T. (1983). Social desirability scales: More substance than style. Journal of Consulting and Clinical Psychology, 51, 882-888. doi:10.1037/0022-006x.51.6.882

Millham, J. (1974). Two components of need for approval score and their relationship to cheating following success and failure. Journal of Research in Personality, 8, 378-392. doi:10.1016/0092-6566(74)90028-2

Musek, J. (2007). A general factor of personality: Evidence for the Big One in the five-factor model. Journal of Research in Personality, 41, 1213-1233. doi:10.1016/j.jrp.2007.02.003

Naemi, B. D., Beal, D. J., & Payne, S. C. (2009). Personality predictors of extreme response style. Journal of Personality, 77, 261-286. doi:10.1111/j.1467-6494.2008.00545.x

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. European Journal of Social Psychology, 15, 263-280. doi:10.1002/ejsp.2420150303

OECD (2014),*TALIS 2013 Results: An International Perspective on Teaching and Learning*, TALIS, OECD Publishing, http://dx.doi.org/10.1787/9789264196261-en.

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. Journal of Applied Psychology, 81, 660-679. doi:10.1037/0021-9010.81.6.660

Paulhus, D. L. (1991). Measurement and control of response biases. In J. Robinson, P. Shaver & L. Wrightsman (Eds.), Measures of personality and social psychological attitudes (Vol. 1, pp. 17-59). San Diego, CA: Academic Press.

Ramanaiah, N. V., Schill, T., & Leung, L. S. (1977). A test of the hypothesis about the two-dimensional nature of the Marlowe-Crowne Social Desirability scale. Journal of Research in Personality, 11, 251-259. doi:10.1016/0092-6566(77)90022-8

Schütz, A. (1998). Assertive, offensive, protective, and defensive styles of self-presentation: A taxonomy. The Journal of Psychology, 132, 611-628. doi:10.1080/00223989809599293

Uziel, L. (2010). Rethinking social desirability scales: From impression management to interpersonally oriented self-control. Perspectives on Psychological Science, 5, 243-262. doi:10.1177/1745691610369465

van Dijk, T. K., Datema, F., Piggen, A.-L. J. H. F., Welten, S. C. M., & van de Vijver, F. J. R. (2009). Acquiescence and extremity in cross-national surveys: Domain dependence and country-level correlates. In A. Gari & K. Mylonas (Eds.), Quod erat demonstrandum: From Herodotus' ethnographic journeys to cross-cultural research. Athens: Pedio Books Publishing.

van Hemert, D. A., van de Vijver, F. J. R., Poortinga, Y. H., & Georgas, J. (2002). Structural and functional equivalence of the Eysenck Personality Questionnaire within and between countries. Personality and Individual Differences, 33, 1229-1249. doi:10.1016/s0191-8869(02)00007-7

Vieluf, S., Kunter, M., & van de Vijver, F. J. R. (2013). Teacher self-efficacy in cross-national perspective. Teaching and Teacher Education, 35, 92-103. doi:10.1016/j.tate.2013.05.006

Yang, Y., Harkness, J. A., Chin, T.-Y., & Villar, A. (2010). Response styles and culture. In J. A. Harkness, M. Broun, B. Edwards, T. P. Johnson, L. Lyberg, P. P. Mohler, B.-E. Pennell & T. W. Smith (Eds.), Survey methods in multinational, multiregional and multicultural contexts (pp. 203-223). New York, NY: Wiley.