DOI: 10.1002/hrm.21852

WILEY

SPECIAL ISSUE ARTICLE

Using multi-item psychometric scales for research and practice in human resource management

Mark A. Robinson

Leeds University Business School, University of Leeds

Correspondence:
Mark A. Robinson, Leeds University Business School, University of Leeds, Leeds, LS2 9JT, UK.
Email: m.robinson@lubs.leeds.ac.uk.

Questionnaires are a widely used research method in human resource management (HRM), and multi-item psychometric scales are the most widely used measures in questionnaires. These scales each have multiple items to measure a construct in a reliable and valid manner. However, using this method effectively involves complex procedures that are frequently misunderstood or unknown. Although there are existing methodological texts addressing this topic, few are exhaustive and they often omit essential practical information. The current article therefore aims to provide a detailed and comprehensive guide to the use of multi-item psychometric scales for HRM research and practice, including their structure, development, use, administration, and data preparation.

KEYWORDS

measurement, multi-item scales, psychometric scales, questionnaires, surveys

1 | INTRODUCTION

Questionnaires are one of the most widely used research methods in the social sciences (Bourque, 2004) and multi-item psychometric scales are the most widely used measures in questionnaires. For instance, 29 of the 62 articles published in *Human Resource Management* during 2015 used multi-item psychometric scales to collect data about topics as diverse as organizational ambidexterity (Halevi, Carmeli, & Brueller, 2015), employee voice (Matsunaga, 2015), and performance management (Festing, Knappert, & Kornau, 2015).

Despite their widespread use, however, the complex principles and procedures underlying multi-item psychometric scales are frequently misunderstood or unknown, even by experienced researchers and practitioners in human resource management (HRM). This is particularly true of HRM practitioners conducting staff surveys (e.g., employee engagement), which ostensibly appear psychometric in nature but often neglect key steps in research design and analysis. Such errors, omissions, and misunderstandings have major implications for HRM research and practice. Unreliable scales prevent the consistent measurement of variables, while scales low in validity may not be measuring the intended variables (Cook, 2009). Such problems can distort research findings, hinder theoretical development, and result in ineffective or even counterproductive HRM practice.

Although methodological guidance is available, most texts focus on specific topics or phases and omit essential practical information. Accordingly, this article addresses all phases of multi-item psychometric scale use—including their structure, development, administration, and the preparation of the collected data for analysis—to provide a comprehensive resource for HRM researchers and practitioners that addresses the many practical issues and common points of confusion. The article will also be useful for researchers and practitioners in other social sciences (e.g., industrial and organizational psychology, management) who use multi-item psychometric scales frequently.

Questionnaires comprise a number of questions that participants are required to answer and are therefore usually a self-report research method (Stone & Turkkan, 2000); although the same methods are sometimes used to rate others, such as supervisor ratings of performance (see, e.g., Yam, Fehr, & Barnes, 2014). Multi-item psychometric scales, the focus of this article, are a specialized type of quantitative measure used in questionnaires (see, e.g., Nevill, Lane, Kilgour, Bowes, & Whyte, 2001) and the most frequently used measure in HRM research. Such scales each have multiple items to measure a variable of interest in a reliable and valid manner (Kline, 2000). Throughout the article, the term *psychometric scale* or simply *scale* is used for brevity rather than the full term *multi-item psychometric scale*. However, it is clear from the literature that several synonymous

terms are used, including *multi-item measures*, *multi-item scales*, *psychometric measures*, and *psychometric scales*. The term *psychometric scale* is preferred here to emphasize the need for reliable and valid measurement.

The article starts by discussing the structure of psychometric scales, then details their use and the various stages of their development, before finally considering their administration and data preparation issues.

2 | STRUCTURE OF PSYCHOMETRIC SCALES

Before proceeding, it is first necessary to define the key components and characteristics of psychometric scales, for there is often confusion here arising from the terminology used. Figure 1 provides an example of a hypothetical psychometric scale measuring managerial support (in lowercase letters) and illustrates its components (in uppercase letters), descriptions of which are provided below (partially adapted from DeVellis, 1991; Kline, 2000).

2.1 | Definition of key terms

Questions/statements. Participants respond to questions or statements about a focal variable.

Response points. A response point is a circle or box in which participants indicate their response to a question or statement, either by ticking, circling (on paper questionnaires), or clicking it (on electronic questionnaires).

Anchors. An anchor is a verbal label accompanying a response point.

Rating scales. A rating scale is the measure along which participants respond to a question or statement. Each question or statement has its own rating scale, comprising a number of response points and accompanying anchors. Usually, for efficiency, a single shared set of anchors is presented for multiple rating scales, as shown in Figure 1.

Items. An item comprises a question or statement about a focal variable and an accompanying rating scale on which participants respond.

Psychometric scales. A psychometric scale comprises multiple items measuring the same focal variable in a reliable and valid manner and yielding parametric data. It should not be confused with a rating scale, despite the shared terminology. A psychometric scale comprises multiple questions or statements, each with their own rating scale. Therefore, a rating scale measures participants' responses to a single question or statement, while a psychometric scale measures participants against a focal variable using multiple items (see DeVellis, 1991). In this article, for brevity and clarity, the single word scale is used to refer to psychometric scales only; rating scales are always referred to by their full name.

When including psychometric scales in questionnaires, researchers have two broad options: they can use existing scales from the published research literature, or they can develop their own scales. Each of these options is discussed in detail below. First, however, the design properties of psychometric scales are discussed.

2.2 | Psychometric scale design properties

Psychometric scales have several design properties that require careful consideration, whether evaluating existing scales or developing new scales, and these are now discussed.

2.2.1 | Likert rating scales

Psychometric scales always use fixed-format rating scales and by far the most widely used are Likert rating scales (Likert, 1932), the focus of this article. Likert rating scales comprise a number of response points, usually 4 to 9, with accompanying verbal anchors. A key feature is that there should be equally appearing intervals (Thurstone, 1929), or identical space perceived by participants, between each response point. This is important because it is a prerequisite of interval or ratio-level data, which in turn is one prerequisite of the parametric data required for psychometric scales (Foster & Parker, 1995).

Such equally appearing intervals (Thurstone, 1929) should be reflected both in the physical presentation of the questionnaire items and also in the meaning of the accompanying verbal anchors. In the former case, the response points should be equidistant from neighboring response points even if this necessitates nonequidistant spacing between verbal anchors of different lengths. In the latter case, antonyms (or opposite terms) should be selected for verbal anchors at equivalent positions on either side of the rating scale, to ensure that there is linguistic symmetry of different valence either side of the rating scale's midpoint. An excellent example of this is the

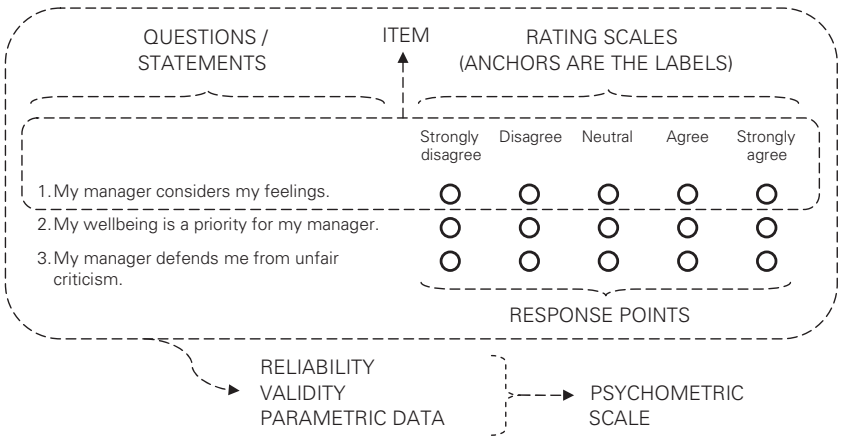

FIGURE 1 Components and characteristics of a psychometric scale
Note: This figure summarizes the more detailed discussions in the section *Structure of Psychometric Scales*, where full references can be found.

commonly used anchor set of *strongly disagree*, *disagree*, *neutral*, *agree*, *strongly agree* (see, e.g., De Jong & Dirks, 2012). Here, the response points on either side of the midpoint have anchors that are exact antonyms—disagree and agree—while the response points two away from the midpoint retain these antonyms but add an identical adverb—strongly. Verbal anchors should ascend from left to right, in level of agreement or level of the rated variable, as this is conventional when listing measurements in written English (e.g., a ruler) and is therefore clearest for participants. Figure 1 illustrates these principles.

Traditionally, each response point has an accompanying verbal anchor, but another common approach is to label only the endpoints of the rating scale. Labeling only the endpoints can alleviate the problem of selecting appropriate labels for each response point, but it increases the difficulty of responding for participants (Darbyshire & McDonald, 2004). Furthermore, although fully labeled response points increase acquiescence, they also reduce extreme responses and increase the clarity of reverse-coded items (Weijters, Cabooter, & Schillewaert, 2010). So, on balance, labeling each response point is usually preferable, unless it is impossible to select balanced and equidistant verbal anchors throughout.

Two modified versions of the traditional Likert rating scale are sometimes used, as described by DeVellis (1991). Semantic differential rating scales present antonymic anchors at either end of a rating scale, along which participants respond. Visual analogue scales replace discrete response points with a continuous line along which participants indicate their response.

2.2.2 | Number of response points

Another key issue with Likert rating scales is deciding how many response points to include. Typically, researchers use between 4 and 9 response points, with some favoring an even number and others an odd number with a midpoint. The optimal number of response points has been frequently debated and examined statistically. For instance, research examining the effect of 2 to 11 response points has shown that reliability, criterion validity, and the ability to discriminate between participants' ratings increase as the number of response points increase, plateauing at around 7 response points (Preston & Colman, 2000). However, other research has indicated that 5-point rating scales yield higher quality data than those with 7 or 11 points (Revilla, Saris, & Krosnick, 2014). Still other research has indicated no difference between data collected from 5-point or 7-point rating scales (Dawes, 2008). Despite these slight disagreements, though, these studies do generally conclude that either 5 or 7 response points yield higher quality data than fewer response points and are more practical than longer rating scales. Thus, researchers should use either 5-point or 7-point rating scales, with decisions between these two options depending on the specifics of the study, such as the variables being investigated, questionnaire space limits, or participant characteristics, as discussed below.

First, for items about some topics where certain responses are more socially desirable—such as positive performance ratings—there can be a tendency for participants to use the corresponding side of the rating scale more frequently (Moorman & Podsakoff, 1992). This leads to highly skewed distributions and effectively a truncated rating scale, with some response points ignored. In these circumstances, a 7-point rating scale will be preferable to a 5-point one, as it will still provide a wide range of responses (see Baumgartner & Steenkamp, 2001, for an extensive review of survey response psychology). For similar reasons, a 7-point rating scale can also be beneficial with participants who are reluctant to select the most extreme responses (Hui & Triandis, 1985). Second, if the questionnaire is to be administered on media with limited display space, such as smartphones, a 5-point rating scale will enable a less cramped presentation than a 7-point one. Finally, when surveying highly educated samples, 7-point rating scales are preferable, as these participants are able to comprehend the additional response complexity, whereas 5-point rating scales are preferable for the general public (Weijters et al., 2010).

There are conflicting views about whether researchers should use an even number of response points with no midpoint or an odd number with a midpoint. Weijters et al. (2010) provide a detailed summary of these counterarguments. Essentially, proponents of the former argue that participants should be forced to choose whether their response to an item is broadly negative or positive. Conversely, proponents of the latter argue that some participants will genuinely have neutral views about some topics, so they should be free to express these accurately. There is some evidence to suggest that the elimination of a midpoint leads to less positive responses to items and may therefore mitigate socially desirable responding (Garland, 1991). Similarly, other evidence suggests that the inclusion of a midpoint increases acquiescence with statements but also decreases extreme ratings (Weijters et al., 2010). Overall, though, there is not yet a clear consensus on this issue, although scales with midpoints are more frequently used than those without.

Finally, when using existing scales, if researchers specifically wish to compare rating levels with those from previous studies, the original rating scales should be used with identical response points and anchors. This may be the case, for instance, if data on norms for different populations exist for particular scales or items.

2.2.3 | Number of scale items

In practice, the number of items in a scale is likely to be predetermined, either by the researchers who published it or during the factor analytic development process, as discussed below. However, scale length is considered here, as it is a key criterion for scale selection.

Conventionally, psychometric scales comprise multiple items. Indeed, it could be argued that this is a prerequisite of psychometric scales, for only multiple items enable the assessment of internal reliability (as detailed later) and reliability is a prerequisite of psychometric scales (Kline, 2000). A minimum of three items per scale is usually recommended, as this number will reliably yield convergent solutions in confirmatory factor analysis (Marsh, Hau, Balla, & Grayson, 1998). However, frequently in research, some problematic items are identified and may therefore have to be deleted, as discussed later. Therefore, it is prudent to include an additional item—so a minimum of four items in a scale—where practical.

The maximum number of items per scale will depend on the complexity of the variable being measured. A larger number of items will be required to capture the richness of multidimensional variables

(see, e.g., Allen & Meyer, 1990). However, this must be balanced against the need for scale brevity to maximize response rates. For this reason, short versions of well-established scales are often developed (see, e.g., Thompson, 2007, for a short version of the mood scale by Watson, Clark, & Tellegen, 1988). If such a short scale is unavailable, researchers may be able to adapt their own from the original publication describing the development of the scale. To do so, only those items with the highest factor loadings for that scale should be selected, and the internal reliability (as detailed later) of the resultant shorter scale should be carefully checked. This is a controversial practice, however, as it can reduce the validity of the scale (Raykov, 2008).

Although still widely recognized as best practice, some researchers have recently questioned the use of multi-item scales. They argue that participants perceive them as repetitive and onerous, therefore reducing response rates, and suggest the use of single-item measures in some circumstances to counter this (see, e.g., Wanous, Reichers, & Hudy, 1997). Accordingly, several single-item measures have been developed that demonstrate good validity against equivalent full-scale versions (see, e.g., Nagy, 2002). Indeed, some researchers believe that for homogeneous construct variables, single-item measures may even be preferable to multi-item scales (Postmes, Haslam, & Jans, 2013), as the latter's specificity may inadvertently exclude key facets of the variable (Scarpello & Campbell, 1983).

3 | USING EXISTING PSYCHOMETRIC SCALES

Several useful repositories of existing psychometric scales are available via the Internet, such as the Academy of Management's (AoM) Measure Chest (n.d.) and the Social-Personality Psychology Questionnaire Instrument Compendium (Reifman, 2014). These provide lists of published scales, categorized by topics. Many researchers also make their own published scales freely available via their personal web pages (see, e.g., Spector, n.d.), so researchers should consult those of influential researchers in their field of interest. However, the primary source of existing scales is peer-reviewed journal articles. Generally, details about the scales used can be found in the *Method* section, often in a subsection entitled *Measures*, with full scales often provided in the appendix.

Finding journal articles with suitable scales can sometimes prove difficult, but has been made considerably easier recently with the introduction of freely available Google Scholar (n.d.) software. Rather than relying solely on keywords like many traditional academic search engines, the advanced search capability of Google Scholar enables researchers to search for exact phrases present anywhere in an article's text, such as the exact name of variables researchers are seeking to measure (e.g., "job satisfaction"). In many cases, this will locate articles reporting empirical studies where data on that variable have been collected using a suitable scale, particularly if the word *questionnaire* or *survey* is used as an additional search term for the article's text. Furthermore, if there are widely recognized leading journals in the field of research interest, it is often worth performing a Google Scholar advanced search restricted to those journals.

Given the many journal articles available about most topics, though, an equally likely problem is that researchers will find several potentially suitable scales that they must choose between. In these cases, and when evaluating existing scales, researchers should use two criteria to select the most appropriate scales: (1) psychometric data, concerning reliability and validity, as extensively detailed in the next section, *Developing Psychometric Scales*, and (2) conceptual fit, as discussed below.

Conceptual fit concerns the extent to which the scale matches the variable that the researcher wishes to measure. Ideally, and in many cases, researchers will be able to find an exact match between scale and variable. In other cases, though, the lack of an exact conceptual fit may necessitate minor modifications to the scale's items. There are no exact rules about acceptable levels of modification, but a useful guideline might be that changing the subject or object of a statement (or question), provided that the statement still relates to the same domain, is generally acceptable provided the researcher checks the psychometric properties (i.e., reliability and validity, see below) of the modified scale against those of the original. So, for instance, Martin, Washburn, Makri, and Gomez-Mejia (2015) modified the subject of the original items from Ryckman, Robbins, Thornton, and Cantrell's (1982) self-efficacy scale (e.g., "I will be able to successfully overcome future challenges") to examine the self-efficacy of firms instead of individuals (e.g., "The firm will be able to successfully overcome future challenges") in their study of CEO risk-taking, as there were no suitable existing scales available. If the scale's items require excessive modification, however, or if there are no conceptually related scales available, which is common in new or emerging research fields, researchers may have to develop their own scales especially for the study. The procedures for doing so are outlined below.

4 | DEVELOPING PSYCHOMETRIC SCALES

If suitable scales do not exist to measure the variables of interest, or if researchers feel that existing scales are inadequate, then they may need to develop their own. The general process for developing a scale is outlined in Figure 2 and described below, drawing on well-established principles discussed in detail in several sources (see, e.g., Hinkin, 1998; Kline, 2000; Matsunaga, 2015, for the development of a scale measuring employee voice strategy). Often, for efficiency, several scales are developed simultaneously using this process (see e.g., Morgeson & Humphrey, 2006). Each stage in Figure 2 will now be described, thereby also providing information for evaluating the psychometric properties of existing scales from the literature.

4.1 | Generate preliminary items

Researchers can use several methods to identify item content, including literature reviews, interviews with experts, and content analysis of existing data sets and resources. This stage is the foundation of the entire process, so it is vital that it is theoretically driven. The nomological network of the variables (Cronbach & Meehl, 1955; Gregory, 2007)—that the generated items will represent—should be

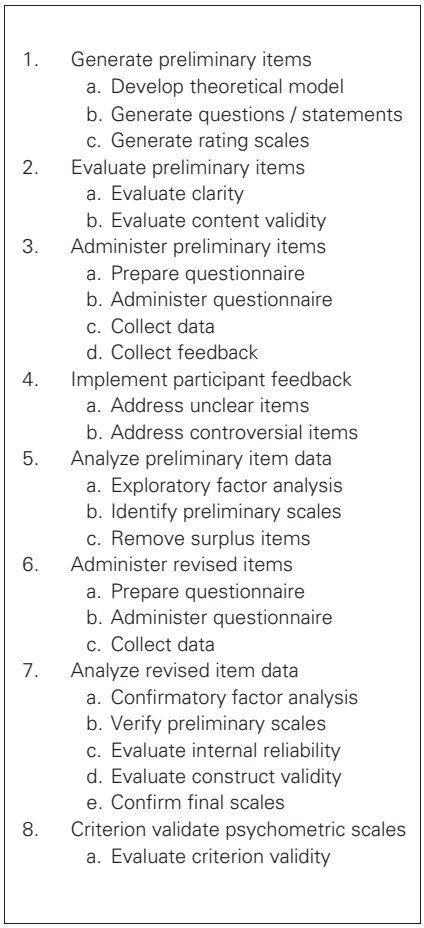

FIGURE 2 Process for developing psychometric scales
Note: This figure summarizes the more detailed discussions in the section *Developing Psychometric Scales*, where full references can be found.

carefully considered, including their theoretically related variables, antecedents, and outcomes. This identifies the unique conceptual territory of the new scales, enabling the themes of the items to be specified and strengthening the construct validity of the resultant scales (see Stage 7). A key consideration here is whether each focal variable (i.e., theoretical construct) is represented by a single dimension (unidimensional) or multiple dimensions (multidimensional), notes Edwards (2001). The latter, he suggests, can be further divided into superordinate constructs (such as personality traits composed of facets) or aggregate constructs (such as different components of job performance). For such multidimensional variables, it is particularly important that the items generated represent the range and richness of the underlying dimensions to ensure sound content validity (see Stage 2).

Questions should be written in a clear and specific manner. Foster and Parker (1995) propose several rules, suggesting that generally questions should:

1. Avoid jargon or specialist terminology, unless well known by the intended participant population.
2. Avoid ambiguity and be specific. For instance, they suggest the term *frequently* is subjective, and it would be better to ask about objective quantities.
3. Avoid combining questions; ask about only one issue at a time.
4. Avoid negatively worded questions, as these can be confusing (although the advice on this issue is mixed, as discussed below).
5. Avoid leading questions, as these can bias participants' responses by suggesting how they should answer.

Furthermore, items should be kept short, where possible, for clarity. Rating scales for the items should be developed in accordance with the design guidelines previously discussed.

Negatively worded items are controversial. They are generally used as "cognitive speed bumps" to prevent participants slipping into inaccurate automatic response patterns (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003, p. 884). However, research has shown that participants generally do not understand such items (Idaszak & Drasgow, 1987), so their benefits are more than outweighed by their costs. Some of this confusion arises from the use of complex double-negatives and similar phrases (Foster & Parker, 1995). In some cases, it may be possible to rephrase the question or statement to alleviate such confusion. For instance, in a scale measuring absenteeism, the statement "I am rarely absent from work" is preferable to the equivalent yet confusing statement "I infrequently do not attend work."

Finally, when developing psychometric scales, researchers should include a larger number of items in the preliminary item pool than the number required for the final psychometric scale(s). The subsequent development process will eliminate many items for statistical or methodological reasons, as discussed below, so some redundancy is useful initially, with double the desired number recommended (Hinkin, 1998).

4.2 | Evaluate preliminary items

Once the items have been generated, they are then evaluated by the researchers for clarity of expression, to ensure that they are easily understandable and assess exactly what is required. Next, the content validity of the items is evaluated; that is the extent to which all facets of the focal variable(s) have been comprehensively addressed by the collective items and without redundancy (Cook, 2009). Often, at this stage, the items are reviewed by a panel of experts, such as those consulted by Sendjaya, Sarros, and Santora (2008) when developing their scales measuring servant leadership. Here, 15 domain experts drawn from academia and business rated each item for relevance, with content validity established when 50% of the experts agreed the item was essential. Similarly, it is possible to calculate a coefficient, the content validity index (CVI), to indicate the proportion of experts who agree about the relevance of single or multiple items, with values exceeding .80 considered acceptable (Polit, Beck, & Owen, 2007).

4.3 | Administer preliminary items

Once evaluated, and improved if necessary, the preliminary items are incorporated into a questionnaire and administered to participants. Often, this stage is referred to as a pilot study, or a pilot questionnaire. This questionnaire should be designed and administered in accordance with the guidance provided elsewhere in this article. However, the pilot questionnaire also usually includes an additional

free-format response section at the end where participants are asked to provide feedback about the items and questionnaire overall.

4.4 | Implement participant feedback

The feedback provided by participants about the preliminary items and questionnaire is analyzed. If a consensus emerges that particular items are unclear or controversial, they should be removed or modified for the next questionnaire (Stage 6). This feedback can also be supplemented with further interviews or focus groups with participants, if required.

4.5 | Analyze preliminary item data

Exploratory factor analysis (EFA) is then conducted on the responses to the preliminary items, to identify the factor structure within the items and thus the preliminary psychometric scales. The basic procedures of EFA are detailed in many statistical textbooks (see, e.g., Tabachnick & Fidell, 2007). The ideal minimum sample size for EFA has been debated frequently, with larger samples and larger item-to-participant ratios considered better (see Osborne & Costello, 2004, for a detailed discussion). Minimum sample sizes of 300 participants are generally advocated (Comrey & Lee, 1992; Tabachnick & Fidell, 2007) unless loadings are particularly high (> .60), in which case 150 participants are adequate (Guadagnoli & Velicer, 1988). A minimum of 10 items per participant are generally recommended (Guadagnoli & Velicer, 1988; Osborne & Costello, 2004). Indeed, medians of 267 participants and 11 items per participant were found in a review of published EFA studies (Henson & Roberts, 2006), corroborating these guidelines.

For unidimensional variables, EFA is performed only on the matrix of correlations between the items, but for multidimensional variables this initial EFA identifies the first-order factors, and a second EFA is then performed on the matrix of correlations between these first-order factors to identify second-order factors (Edwards, 2001; Gorsuch, 1983). So, when used for psychometric scale development, first-order factors would identify subscales that are nested within the wider construct represented by the second-order factor.

Consideration should be given here to the optimal number of items in each scale, with factor loadings examined accordingly. Comrey and Lee (1992) have proposed statistical loading thresholds to aid factor interpretability, and these could also be usefully applied to determine how many items to retain in each psychometric scale. Noting how squared factor loadings indicate the proportion of shared variance between those items and the factors on which they load, they suggested that items loading over .71 (50% shared variance) have an excellent fit with the factor, those loading over .63 (40% shared variance) a very good fit, over .55 (30% shared variance) a good fit, and over .45 a fair fit (20% shared variance). This resonates with Costello and Osborne's (2005) more recent recommended threshold for loadings of .50 or higher. Furthermore, both pairs of authors caution against retaining items loading below .30. So items with factor loadings above this .45 threshold would therefore make excellent scales when combined.

Provided that such statistical criteria are satisfied, however, the choice of how many items to select for each scale is likely to be determined by practical considerations. Researchers should strive to achieve an optimal balance between parsimony and comprehensive theoretical coverage of the focal variables, removing further surplus preliminary items where required.

4.6 | Administer revised items

Once the preliminary items have been analyzed, modified as necessary, and reduced to an optimal number per scale, these revised items are then administered to a further sample of participants in another questionnaire. Again, this questionnaire should be designed and administered in accordance with the guidance provided throughout this article. Here, researchers should also administer existing scales measuring conceptually related and unrelated variables to enable construct validity to be assessed (see, e.g., Lewis, 2003), as detailed in the description of Stage 7 below.

4.7 | Analyze revised item data

Confirmatory factor analysis (CFA) is then conducted on the responses to the revised items to verify the factor structure identified by the initial EFA. If verified, the item composition of the scale(s) can be considered finalized. If the CFA does not support the factor structure identified by the initial EFA, however, then it may be necessary to readminister the revised items to a further sample of participants until statistical consensus is achieved regarding factor structure. The sample size recommendations provided for EFA above are generally also of relevance to CFA (Mundfrom, Shaw, & Ke, 2005; Tabachnick & Fidell, 2007). Specifically, though, a minimum sample of 200 participants has frequently been recommended for CFA (Barrett, 2007; Tabachnick & Fidell, 2007). Once the final factor structure has been confirmed, the internal reliability of the scales should then be assessed.

Psychometric scales use multiple items measuring the same focal variable so that the consistency or internal reliability with which participants respond to them can be assessed. Reliability is a prerequisite for validity, and both are essential characteristics of psychometric scales (Kline, 2000). Cronbach's alpha coefficient (α; Cronbach, 1951) is the most frequently used statistic for this purpose. There is widespread debate about what the minimum acceptable alpha coefficient level is for psychometric scales. Traditionally, the figure of $\alpha \geq .70$ was widely suggested, although the origin was uncertain. Cortina (1993) discusses alpha in considerable depth, noting that the same mean inter-item correlation will yield higher alpha coefficients for longer scales than shorter scales, and advises cautious interpretation. Nevertheless, he suggests $\alpha \geq .75$ as the conventional accepted level.

If researchers discover the alpha coefficient of a scale they have used is below .75, they may wish to consider deleting an item to increase this coefficient. Statistical analysis software such as SPSS (n.d.) can calculate alpha coefficients with each item deleted alongside the alpha coefficient for the overall scale, helping to identify rogue items for potential deletion. However, such item deletion is a controversial practice, with some arguing that it dilutes the conceptual coverage and validity of the scale (Raykov, 2008). If the rogue

item is a negatively worded one, however, the problem is likely a methodological artifact (see Idaszak & Drasgow, 1987), as discussed above, and it should therefore be deleted.

However, for CFA, researchers are now increasingly using the composite reliability method of Dillon-Goldstein's rho (or Jöreskog's rho) (ρ_c), for which values over .70 indicate acceptable reliability (Chin, 1998; Werts, Linn, & Jöreskog, 1974). Unlike Cronbach's alpha, it does not assume each of a scale's constituent items is of equal importance, as it is based on the factor loadings instead of the correlations between items, and is therefore more accurate (Esposito Vinzi, Trinchera, & Amato, 2010).

Finally in this stage, construct validity is assessed, which concerns whether the scales measure the constructs they claim to (Cook, 2009). This should be demonstrated in three ways. First, the factor analyses performed in Stage 5 and here in Stage 7 will establish some construct validity through the distinct factor structures identified and the absence of cross-loadings (Tabachnick & Fidell, 2007). Second, to establish convergent construct validity, scale scores for each variable should be highly correlated (i.e., converge) with scores from other established measures of the same variable and also with measures of other variables from within that variable's nomological network of theoretically related constructs (Cronbach & Meehl, 1955; Gregory, 2007), administered in Stage 6. Third, to establish divergent (or discriminant) construct validity, scale scores for each variable should be uncorrelated (i.e., diverge) with scores from theoretically unrelated constructs (Gregory, 2007), also administered in Stage 6. For instance, Maynes and Podsakoff (2014) established the construct validity of their four measures of employee voice behaviors through examining the relationships between these scale scores and data from related and unrelated measures of the Big Five personality traits. As a guide, convergent construct validity would be demonstrated by medium to large correlations exceeding .30, while small correlations below .20 would indicate divergent construct validity (Cohen, 1992; see also Lewis, 2003).

4.8 | Criterion validate psychometric scales

By this stage, the reliability of the scales has been established, as has their content validity through the generation and evaluation of the items, and their construct validity through the two factor analyses and the convergent and divergent analyses. So the aim of this stage is to establish the criterion validity of the scale(s). Here, participants' scores on the scale(s) from Stage 7 are statistically correlated with independent objective data measuring the same or related variables for each participant (Cook, 2009). For instance, a self-report scale measuring staff absenteeism could be criterion validated against absence data from official organizational records, using suitable identification codes to match the two data sources. The criterion data against which scale scores are validated can either be collected at the same time the scales are completed, to establish concurrent criterion validity, or at a future date, to establish predictive criterion validity (Cook, 2009). Drawing on Cohen's (1992) guidance about effect sizes, correlations exceeding .30 would indicate reasonable criterion validity, with correlations exceeding .50 being excellent. Resonating with these values, a minimum correlation threshold of .40 is therefore recommended for establishing sound criterion validity (Peers, 1996).

Following this final validation, the psychometric scales are now ready to use for research purposes. They may also be published, with accompanying psychometric data concerning their reliability and validity, for use by other researchers.

Finally, sometimes at this stage the scales are readministered to the same participants from Stage 6 and their scores are compared to establish test–retest reliability (Cook, 2009). Pearson correlations exceeding .80 would indicate acceptable test–retest reliability, but this is only a relevant concept for the few variables expected to exhibit stability over time such as personality (Kline, 2000).

Table 1 provides a summary of the different types of reliability and validity of relevance to developing psychometric scales, as discussed above.

5 | ADMINISTRATION

Psychometric scales are administered within questionnaires, so when using them for research there are several practical issues to consider, and these are now reviewed. The first broad issue is the content of the remainder of the questionnaire, including its introduction, the generation of identification codes, and demographic questions. The second broad issue concerns the format of questionnaire administration and the implications of this for the presentation of the psychometric scales to participants.

5.1 | Introductory information

First, participants are briefly provided with an overview of the research, the purpose of their involvement, and the contact details of the researchers. This engages participants, and also provides sufficient information for them to give their informed consent to participate (American Psychological Association [APA], 2010). Typically, the research overview is relatively general, justifying their participation but not detailing specific research questions or hypotheses. Indeed, excessive information of this nature may prime participants to respond in particular ways, which may bias the research.

Second, a number of statements relating to research ethics are presented. Conventions can vary by discipline, but the APA (2010) provides extensive guidelines, the key principles of which are summarized below. First, participants are told that their involvement is entirely voluntary and that they have the right to withdraw at any time. They are also assured that any information they provide will remain entirely confidential, and that results will only be presented in an aggregated format so that no individual's responses are identifiable. Then, they are asked to provide their informed consent to participate, either by answering a direct question to this effect or by being informed that their continuation implies this. Most institutions and organizations in which researchers work will have their own formal ethical clearance procedures based on similar principles. Finally, questions about sensitive or controversial topics will usually require thorough justification via an ethics committee.

TABLE 1 Types of reliability and validity relevant to psychometric scale development

Concept	Definition	Measure
Internal reliability	All the items comprising the psychometric scale are measuring the same variable consistently.	Cronbach's alpha (α) ≥ .75
		Dillon-Goldstein's rho (or Jöreskog's rho) (ρ_c) ≥ .70
Test–retest reliability	Scores on the psychometric scale are consistent over time for the same person (i.e., scores attained at different times).	Pearson correlation (r) ≥ .80
	Test–retest reliability is only relevant for variables expected to be stable over time (e.g., personality).	
Content validity	Collectively, the items comprising the psychometric scale address all key aspects of the variable and no irrelevant aspects.	Content validity index (CVI) ≥ .80
Construct validity	The psychometric scale is measuring the specific variable it claims to measure (and not another similar variable).	Items load onto a single factor that is distinct from other factors (i.e., no cross-loadings)
	Convergent construct validity	
	Scores on the psychometric scale are strongly related to scores on psychometric scales measuring conceptually related variables.	Pearson correlation (r) ≥ .30
	Divergent (or discriminant) construct validity	
	Scores on the psychometric scale are not strongly related to scores on psychometric scales measuring conceptually unrelated variables.	Pearson correlation (r) < .20
Criterion validity	Scores on the psychometric scale are strongly related to objective measures of the same variable.	Pearson correlation (r) ≥ .40
	Concurrent criterion validity	
	Scores on the psychometric scale are strongly related to objective measures of the same variable measured at the same time.	
	Predictive criterion validity	
	Scores on the psychometric scale are strongly related to objective measures of the same variable measured at a future time.	

Note: This table summarizes the more detailed discussions in the section *Developing Psychometric Scales*, where full references can be found.

5.2 | Identification codes

In many research studies, questionnaires can be completed entirely anonymously, so no identification features are required. However, in some instances, it is necessary for researchers to be able to identify participants. For example, participants may be distributed between various groups (e.g., teams or organizations) and researchers may wish to analyze the data at a group level, such as Halevi et al.'s (2015) study of organizational ambidexterity in strategic business units. In such instances, this should be explained to participants in the introductory section and questionnaires should carry a suitable code (e.g., Team 1) to enable their grouping when returned.

A second example is the use of longitudinal research designs, where data are collected from participants—using questionnaires and/or other methods—at two or more time points, requiring researchers to match data from the same participant. For example, Sturges and Guest (2004) examined work–life balance in recent graduates, tracking them from before they started work and into their first appointment, by administering questionnaires at three time points six months apart. Generally, participants are unwilling to divulge readily identifiable personal information (e.g., names), however, so they should each be asked to generate an anonymous and unique identification code to include on each questionnaire they complete. This code should contain information that will not change, such as a nine-letter code comprising the first three letters from each of the following three words: (1) first pet's name, (2) hometown, (3) favorite sports team. In this way, the code does not have to be remembered, although this is desirable, as it can be generated again through asking the same questions. Then, each time participants complete a new questionnaire, they should be asked to provide this code to enable matching with their previous questionnaires.

5.3 | Demographic questions

Often in research, it is necessary to collect demographic data from participants. Sometimes, this is an integral part of the research itself; for instance, when conducting research examining age or gender (see e.g., Festing et al., 2015, for a study of performance management preferences between genders). In other instances, it is necessary to control for demographic variables when performing statistical analyses (see e.g., Mäkelä, Kinnunen, & Suutari, 2015). When conducting research in organizations, it may also be desirable to collect data concerning variables such as participants' roles, seniority, and experience. For standardization, demographic data are generally best collected using questions with fixed response categories, from which participants select the appropriate response. One possible exception is questions concerning time, such as age and organizational tenure, where the collection of exact data (e.g., 33 years)—provided this does

not identify participants—can be subsequently coded into fixed categories (e.g., 30–34 years) if required.

5.4 | Administration format: Paper or electronic

Generally, there are two broad approaches for administering questionnaires: paper-based and electronic. In recent years, due to the advent of specialist software, the latter approach is increasingly used; however, each has its advantages and disadvantages.

The response rates and financial costs of each approach were systematically investigated by Greenlaw and Brown-Welty (2009). Questionnaires administered electronically through the Internet resulted in a higher response rate (52.46%) and a substantially lower cost per response ($0.64) than those administered in a paper format (42.03%, $4.78). Although providing participants with both options yielded the highest response rate of all (60.27%), the high cost of doing so outweighed this advantage ($3.61). Overall, then, the electronic Internet-based approach was superior. Administering questionnaires via Internet-based services (e.g., Qualtrics, n.d.) is also extremely efficient, as it enables data to be downloaded electronically, which greatly reduces the time taken to enter and format data for statistical analysis. Given the global reach of the Internet, it is also possible to recruit large numbers of participants with relative ease.

Increasingly, electronic questionnaires are being administered on handheld digital devices such as PDAs, smartphones, and tablets, and this approach has great research potential (Miller, 2012). In particular, the portability and convenience of such devices greatly facilitates the repeated collection of diary data over extended periods (Robinson, 2012). Reassuringly, research indicates that data collected from identical surveys on computers and smartphones are comparable (de Bruijne & Wijnant, 2013). However, items and responses should be clearly formatted for display on such small screens, in accordance with the guidelines further below.

Despite these advantages, though, there are still two circumstances where paper-based questionnaire administration may be preferable. First, some participants may not have access to the Internet (e.g., factory workers), so paper questionnaires are the only practical option. Second, in some cases, paper questionnaires may be more convenient, particularly if potential participants are gathered together in a single venue (such as a conference) and have some spare time to participate.

5.5 | Questionnaire presentation: Methodological issues

There are three methodological issues concerning the presentation of items in questionnaires that researchers should pay careful attention to, as discussed below. In principle, these issues relate to questionnaires administered in any format; although, in practice, they relate mainly to electronic questionnaires.

First, if the rating scale anchors are only provided at the start of the questionnaire, participants may lose sight of them as they scroll down the screen, potentially leading to confusion or incorrect responses. This issue can be partly resolved by displaying the questionnaire on a number of shorter consecutive screens, each with the anchors displayed at the top, to replicate a conventional paper questionnaire format. A further option is to display the scale anchors at both the top and bottom of each screen, or above shorter blocks of items, so that one set of anchors is always visible to participants.

Second, there are numerous different orders in which questionnaire items and scales can be displayed, and these may have subtle effects on how participants respond. In general, it is best to cover important topics earlier and sensitive topics later. In this way, if participants do not complete the questionnaire, due to length or sensitivity, some useful data may still be collected.

Third, in recent years, researchers have debated whether questionnaire items should be grouped by topic, or presented in a mixed or random order. Podsakoff et al. (2003) summarize the key issues in this debate, as discussed below. Essentially, advocates of the former approach argue that grouping items is clearer for participants and allows them to carefully consider each topic holistically. However, critics of this approach argue that by grouping multiple similar items, common method bias may lead to artificially high consistency between responses to a scale's items. Indeed, Podsakoff et al. note that clustering items from the same scales together inflates intra-scale correlations, and thus also inflates Cronbach's alpha internal reliability, while mixing them inflates inter-scale correlations, and thus some bias is inevitable either way. They therefore conclude that the issue has yet to be resolved.

Given the lack of consensus, then, the simplest approach is probably best. Unless the process of mixed or random item ordering can be fully electronically automated, the potential for subsequent confusion and mistakes—when regrouping the items for analysis—would suggest that the simpler method of grouping the items by scale (or theme) is the best procedure when administering questionnaires.

6 | DATA PREPARATION

6.1 | Numerically coding responses

Once researchers receive the data, the first step in calculating scale scores is to numerically code the response points to quantify the participants' response data. Consecutive ascending whole numbers are almost always used to reflect the equally appearing intervals (Thurstone, 1929) indicated by the verbal anchors. This is the recommended option. So, for instance, *strongly disagree* could be coded 1, *disagree* coded 2, *neutral* coded 3, *agree* coded 4, and *strongly agree* coded 5 (see e.g., Albirini, 2006).

Through convention, most researchers code the lowest response point as 1 rather than 0 (see, e.g., Albirini, 2006), although some do the latter (see e.g., Bolger, Zuckerman, & Kessler, 2000). Either way, inter-scale correlations will be the same; although scale scores (see below) will naturally be 1 higher in the former case. However, there are good reasons for coding response points from 0 upwards rather than from 1. First, if scale scores are displayed in a graph, then it is possible to start the *y*-axis at 0, which makes intuitive sense and is the default option in most graphics software. Indeed, a very common error is to display scale scores measured using 1–5 coding on graphs with 0–5 *y*-axes, thereby erroneously inflating perceived scores.

Second, and related, when scale scores starting at 1 are reported as having a maximum score of 5, for example, or alternatively as measured on a 5-point scale, a misperception often arises where readers erroneously assume that 2.5 is the midpoint rather than the true value of 3, again serving to erroneously inflate perceived scores.

Finally, where negatively worded items have been used, these need to be reverse-coded before proceeding. So, using a traditional 5-point rating scale, the reverse coding would proceed as follows: *strongly disagree* (1 → 5), *disagree* (2 → 4), *neutral* (3 → 3), *agree* (4 → 2), *strongly agree* (5 → 1). To check the accuracy of this recoding, it is prudent to correlate the original negatively worded items with their reverse-coded counterpart items, to ensure that correlations are $r = -1.00$ as they should be. Finally, it is important to note that negatively worded items are identified relative to the variable the scale is measuring. So, for example, an item about alertness would be a negatively worded one in a scale measuring fatigue, even if the item itself does not contain a negative prefix (e.g., *not* or *un*-).

6.2 | Calculating scale scores

There are two ways in which scale scores can be calculated. First, the mean rating of all items comprising the scale can be calculated. Second, the ratings of all items comprising the scale can be aggregated. If there are no missing data, either option will yield identical inter-scale correlations, albeit with different scale scores, naturally. However, in reality, there are almost always missing data, in which case aggregating item ratings will yield lower scale scores than appropriate for participants with missing data. Consequently, the first option—calculating the mean rating of all items comprising the scale—is strongly recommended, and this method is almost always used in published research. A further advantage of this mean item rating method is that the scale score is calibrated to the original rating scale—for instance, a scale score of 3.4 on a 1–5 rating scale—and therefore has more meaning for readers than an aggregated item scale score that is more reflective of the number of items than the strength of response.

When calculating the scale score from the mean of its constituent items' ratings, however, it is necessary to decide how many of the scale's items a participant must respond to for this calculation to be valid. Graham (2009) suggests that participants must have responded to at least 50% of a scale's constituent items before their scale scores can be calculated using the mean item rating method. He also cautions that the scale's Cronbach alpha internal reliability should be high and the items answered should adequately represent the scale's construct. While agreeing with these latter two restrictions, Newman (2014) suggests that even one item is sufficient to calculate a scale score, reasoning that this is less wasteful of precious data and therefore increases statistical power. To balance these competing demands, a conservative rule of thumb might therefore be that: (a) a threshold of responses to at least 50% of a scale's items should be reached before calculating a scale score, unless (b) this approach reduces the sample size to below recommended levels for statistical analyses, in which case scale scores should be calculated from responses to one or more items provided the Cronbach's alpha internal reliability of the scale is high ($\alpha \geq .75$; Cortina, 1993). Whichever approach is used, SPSS (n.d.) software offers researchers an option to specify the minimum number of items for which a response must have been recorded before a scale score is calculated.

There are notable exceptions to this recommendation, however. Some scales have aggregated item scale scores that correspond to thresholds or particular critical levels of a variable. For instance, the revised Negative Acts Questionnaire has threshold scores corresponding to different degrees of workplace bullying (Notelaers & Einarsen, 2013). In these cases, an aggregated item scale score may be required, and precautions should therefore be taken to ensure participants respond to all items.

7 | CONCLUSION

Psychometric scales are arguably the most frequently used research method in the social sciences. However, their effective development and use requires a detailed knowledge of technical procedures and issues that are frequently not well taught or misunderstood. It is hoped that this article will therefore provide HRM researchers and practitioners with a solid grounding in this important method for the benefit of future research and practice.

ORCID

Mark A. Robinson http://orcid.org/0000-0001-5535-8737

REFERENCES

Academy of Management (AoM). (n.d.). *Measure chest*. Research Methods Division, AoM. Retrieved from http://rmdiv.org/?page_id=104

Albirini, A. (2006). Teachers' attitudes toward information and communication technologies: The case of Syrian EFL teachers. *Computers & Education*, *47*(4), 373–398.

Allen, N. J., & Meyer, J. P. (1990). The measurement and antecedents of affective, continuance and normative commitment to the organization. *Journal of Occupational Psychology*, *63*(1), 1–18.

American Psychological Association (APA). (2010). Ethical principles of psychologists and code of conduct. Retrieved from http://www.apa.org/ethics/code/principles.pdf

Barrett, P. (2007). Structural equation modeling: Adjudging model fit. *Personality and Individual Differences*, *42*(5), 815–824.

Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, *38*(2), 143–156.

Bolger, N., Zuckerman, A., & Kessler, R. C. (2000). Invisible support and adjustment to stress. *Journal of Personality and Social Psychology*, *79*(6), 953–961.

Bourque, L. B. (2004). Self-administered questionnaire. In M. S. Lewis-Beck, A. Bryman, & T. Futing Liao (Eds.), *The Sage encyclopedia of social science research methods* (Vol. *3*, pp. 1012–1013). London, England: Sage.

Chin, W. W. (1998). The partial least squares approach to structural equation modeling. In G. A. Marcoulides (Ed.), *Modern methods for business research* (pp. 295–336). Mahwah, NJ: Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159.

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cook, M. (2009). *Personnel selection: Adding value through people* (5th ed.). Chichester, England: Wiley-Blackwell.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98–104.

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analyses. *Practical Assessment, Research & Evaluation*, *10*(7), 1–9.

Cronbach, L. J. (1951). Coefficient alpha and the internal consistency of tests. *Psychometrika*, *16*(3), 297–334.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302.

Darbyshire, P., & McDonald, H. (2004). Choosing response scale labels and length: Guidance for researchers and clients. *Australasian Journal of Market Research*, *12*(2), 17–26.

Dawes, J. (2008). Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. *International Journal of Market Research*, *50*(1), 61–77.

de Bruijne, M., & Wijnant, A. (2013). Comparing survey results obtained via mobile devices and computers: An experiment with a mobile web survey on a heterogeneous group of mobile devices versus a computer-assisted web survey. *Social Science Computer Review*, *31*(4), 482–504.

De Jong, B. A., & Dirks, K. T. (2012). Beyond shared perceptions of trust and monitoring in teams: Implications of asymmetry and dissensus. *Journal of Applied Psychology*, *97*(2), 391–406.

DeVellis, R. F. (1991). *Scale development: Theory and applications*. London, England: Sage.

Edwards, J. R. (2001). Multidimensional constructs in organizational behavior research: An integrative analytical framework. *Organizational Research Methods*, *4*(2), 144–192.

Esposito Vinzi, V., Trinchera, L., & Amato, S. (2010). PLS path modeling: From foundations to recent developments and open issues for model assessment and improvement. In V. Esposito Vinzi, W. W. Chin, J. Henseler, & H. Wang (Eds.), *Handbook of partial least squares* (pp. 47–82). Berlin, Germany: Springer-Verlag.

Festing, M., Knappert, L., & Kornau, A. (2015). Gender-specific preferences in global performance management: An empirical study of male and female managers in a multinational context. *Human Resource Management*, *54*(1), 55–79.

Foster, J. J., & Parker, I. (1995). *Carrying out investigations in psychology: Methods and statistics*. Leicester, England: British Psychological Society (BPS) Books.

Garland, R. (1991). The mid-point on a rating-scale: Is it desirable? *Marketing Bulletin*, *2*, 66–70.

Google Scholar. (n.d.). Academic literature search software. http://scholar.google.com

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549–576.

Greenlaw, C., & Brown-Welty, S. (2009). Testing assumptions of survey mode and response cost: A comparison of web-based and paper-based survey methods. *Evaluation Review*, *33*(5), 464–480.

Gregory, R. J. (2007). *Psychological testing: History, principles, and applications*. London, England: Pearson Education.

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin*, *103*(2), 265–275.

Halevi, M. Y., Carmeli, A., & Brueller, N. N. (2015). Ambidexterity in SBUs: TMT behavioral integration and environmental dynamism. *Human Resource Management*, *54*(Suppl. 1), 223–238.

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement*, *66*(3), 393–416.

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods*, *1*(1), 104–121.

Hui, C. H., & Triandis, H. C. (1985). The instability of response sets. *Public Opinion Quarterly*, *49*(2), 253–260.

Idaszak, J. R., & Drasgow, F. (1987). A revision of the job diagnostic survey: Elimination of a measurement artifact. *Journal of Applied Psychology*, *72*(1), 69–74.

Kline, P. (2000). *Handbook of psychological testing* (2nd ed.). London, England: Routledge.

Lewis, K. (2003). Measuring transactive memory systems in the field: Scale development and validation. *Journal of Applied Psychology*, *88*(4), 587–604.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *22*(140), 5–53.

Mäkelä, L., Kinnunen, U., & Suutari, V. (2015). Work-to-life conflict and enrichment among international business travelers: The role of international career orientation. *Human Resource Management*, *54*(3), 517–531.

Marsh, H. W., Hau, K. T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research*, *33*(2), 181–220.

Martin, G., Washburn, N., Makri, M., & Gomez-Mejia, L. R. (2015). Not all risk taking is born equal: The behavioral agency model and CEO's perception of firm efficacy. *Human Resource Management*, *54*(3), 483–498.

Matsunaga, M. (2015). Development and validation of an employee voice strategy scale through four studies in Japan. *Human Resource Management*, *54*(4), 653–671.

Maynes, T. D., & Podsakoff, P. M. (2014). Speaking more broadly: An examination of the nature, antecedents, and consequences of an expanded set of employee voice behaviors. *Journal of Applied Psychology*, *99*(1), 87–112.

Miller, G. (2012). The smartphone psychology manifesto. *Perspectives on Psychological Science*, *7*(3), 221–237.

Moorman, R. H., & Podsakoff, P. M. (1992). A meta-analytic review and empirical test of the potential confounding effects of social desirability response sets in organizational behaviour research. *Journal of Occupational and Organizational Psychology*, *65*(2), 131–149.

Morgeson, F. P., & Humphrey, S. E. (2006). The work design questionnaire (WDQ): Developing and validating a comprehensive measure for assessing job design and the nature of work. *Journal of Applied Psychology*, *91*(6), 1321–1339.

Mundfrom, D. J., Shaw, D. G., & Ke, T. L. (2005). Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing*, *5*(2), 159–168.

Nagy, M. S. (2002). Using a single-item approach to measure facet job satisfaction. *Journal of Occupational and Organizational Psychology*, *75*(1), 77–86.

Nevill, A. M., Lane, A. M., Kilgour, L. J., Bowes, N., & Whyte, G. P. (2001). Stability of psychometric questionnaires. *Journal of Sports Sciences*, *19*(4), 273–278.

Newman, D. A. (2014). Missing data: Five practical guidelines. *Organizational Research Methods*, *17*(4), 372–411.

Notelaers, G., & Einarsen, S. (2013). The world turns at 33 and 45: Defining simple cutoff scores for the Negative Acts Questionnaire—Revised in a representative sample. *European Journal of Work and Organizational Psychology*, *22*(6), 670–682.

Osborne, J. W., & Costello, A. B. (2004). Sample size and subject to item ratio in principal components analysis. *Practical Assessment, Research & Evaluation*, *9*(11), 1–9.

Peers, I. S. (1996). *Statistical analysis for education and psychology researchers*. London, England: Falmer Press.

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879–903.

Polit, D. F., Beck, C. T., & Owen, S. V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health*, *30*(4), 459–467.

Postmes, T., Haslam, S. A., & Jans, L. (2013). A single-item measure of social identification: Reliability, validity, and utility. *British Journal of Social Psychology*, *52*(4), 597–617.

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminatory power, and respondent preferences. *Acta Psychologica*, *104*, 1–15.

Qualtrics. (n.d.). Questionnaire administration software. http://www.qualtrics.com/

Raykov, T. (2008). Alpha if item deleted: A note on loss of criterion validity in scale development if maximizing coefficient alpha. *British Journal of Mathematical and Statistical Psychology*, *61*, 275–285.

Reifman, A. (2014). Social-personality psychology questionnaire instrument compendium. http://www.webpages.ttu.edu/areifman/qic.htm

Revilla, M. A., Saris, W. E., & Krosnick, J. A. (2014). Choosing the number of categories in agree–disagree scales. *Sociological Methods & Research*, *43*(1), 73–97.

Robinson, M. A. (2012). How design engineers spend their time: Job content and task satisfaction. *Design Studies*, *33*(4), 391–425.

Ryckman, R. M., Robbins, M. A., Thornton, B., & Cantrell, P. (1982). Development and validation of a physical self-efficacy scale. *Journal of Personality and Social Psychology*, *42*(5), 891–900.

Scarpello, V., & Campbell, J. P. (1983). Job satisfaction: Are all the parts there? *Personnel Psychology*, *36*(3), 577–600.

Sendjaya, S., Sarros, J. C., & Santora, J. C. (2008). Defining and measuring servant leadership behaviour in organizations. *Journal of Management Studies*, *45*(2), 402–424.

Spector, P. E. (n.d.). Psychological instrument resources. http://shell.cas.usf.edu/~pspector/scalepage.html

SPSS. (n.d.). Statistical analysis software. http://www-01.ibm.com/software/analytics/spss/

Stone, A. A., & Turkkan, J. S. (2000). Preface. In A. A. Stone, J. S. Turkkan, C. A. Bachrach, J. B. Jobe, H. S. Kurtzman, & V. S. Cain (Eds.), *The science of self-report* (pp. ix–xi). Mahwah, NJ: Erlbaum.

Sturges, J., & Guest, D. (2004). Working to live or living to work? Work/life balance early in the career. *Human Resource Management Journal*, *14*(4), 5–20.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). London, England: Pearson/Allyn & Bacon.

Thompson, E. R. (2007). Development and validation of an internationally reliable short-form of the Positive and Negative Affect Schedule (PANAS). *Journal of Cross-Cultural Psychology*, *38*(2), 227–242.

Thurstone, L. L. (1929). Theory of attitude measurement. *Journal of Experimental Psychology*, *12*(3), 214–224.

Wanous, J. P., Reichers, A. E., & Hudy, M. J. (1997). Overall job satisfaction: How good are single-item measures? *Journal of Applied Psychology*, *82*(2), 247–252.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070.

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, *27*(3), 236–247.

Werts, C. E., Linn, R. L., & Jöreskog, K. G. (1974). Intraclass reliability estimates: Testing structural assumptions. *Educational and Psychological Measurement*, *34*(1), 25–33.

Yam, K. C., Fehr, R., & Barnes, C. M. (2014). Morning employees are perceived as better employees: Employees' start times influence supervisor performance ratings. *Journal of Applied Psychology*, *99*(6), 1288–1299.

AUTHOR'S BIOGRAPHY

Mark Robinson holds a PhD in Organizational Psychology from the University of Leeds, where he currently works as a faculty member in Leeds University Business School. He is also Deputy Director of the Socio-Technical Centre, an interdisciplinary research centre, and a member of the Workplace Behaviour Research Centre. His research interests include human performance, group behavior, social cognition, and complex systems.

How to cite this article: Robinson MA. Using multi-item psychometric scales for research and practice in human resource management. *Hum Resour Manage.* 2018;57:739–750. https://doi.org/10.1002/hrm.21852