Improving scale adaptation practices in information systems research: Development and validation of a cognitive validity assessment method

Jean-charles Pillet, Kevin D Carillo, Claudio Vitari, Federico Pigni

Received: 21 January 2021 | Revised: 9 December 2022 | Accepted: 23 December 2022

DOI: 10.1111/isj.12428

FOCUS ON RESEARCH METHODS

WILEY

Improving scale adaptation practices in information systems research: Development and validation of a cognitive validity assessment method

Jean-Charles Pillet[1] | Kevin D. Carillo[1] | Claudio Vitari[2] | Federico Pigni[3]

[1]Department of Information Management, TBS Business School, Toulouse, France

[2]Aix Marseille Univ, CERGAM, FEG, Aix-en-Provence, France

[3]Department of Management, Technology & Strategy, Grenoble Ecole de Management, Grenoble, France

Correspondence

Jean-Charles Pillet, Department of Information Management, TBS Business School, 1 Place Alphonse Jourdain, 31000 Toulouse, France.
Email: jean-charles.pillet@tbs-education.fr

Abstract

Scale adaptation, where authors alter the wording of an already published scale, is a deeply rooted social practice in IS research. This paper argues that the time is ripe to question this activity as well as the beliefs that have progressively formed around it. We identify and challenge five fallacious scale adaptation beliefs that hinder the development of more robust measure development norms. Contributing to this area of research, this paper offers a conceptual definition of the cognitive validity concept, defined as the extent to which a scale is free of problematic item characteristics (PICs) that bias the survey response process and subsequent empirical results. Building on this conceptualization effort, a new methodological process for assessing the cognitive validity of adapted IS measures is introduced. Through a series of three programmatic studies, we find converging evidence that the method can benefit the IS field by making the scale adaptation process more robust, transparent, and consistent. Along with the method, we introduce a new index that IS scholars can use to benchmark the cognitive quality of their scales against venerable IS measures. We discuss the implications of our work for IS research (including detailed implementation guidelines) and provide directions for future research on measurement in IS.

KEYWORDS
measure development, scale adaptation, survey research, survey response model

1 | INTRODUCTION

Adapting self-reported measures is a common practice in empirical information systems (IS) research due to the high domain specificity of IT-enabled organisational and societal phenomena. Scale adaptation is an umbrella term used to indicate that an author changed something in a measurement scale (Heggestad et al., 2019): the number of items, the target of the scale (e.g., individual, team, organisation), the situational context of the items, the format of the scale, the number of response options, and so on. This paper focuses on scale adaptations that involve purposeful item wording modifications, a largely opaque practice that currently lacks explicit standards and norms.

Item wording modifications can potentially undermine a scale's cognitive validity, a broad concept that refers to the ability of a scale to generate unbiased responses. Formally, cognitive validity is defined as the extent to which a scale's items are free of problematic item characteristics (PICs) that bias the response process and subsequent empirical results (Podsakoff et al., 2003; Schwarz, 1999; Sudman et al., 2010; Tourangeau et al., 2000). Well-known examples of PICs are double-barreledness, ambiguity, leading, socially desirable wording, and so on. When respondents' reaction to a scale's items is driven by PICs rather than by the substantive meaning of the items, the cognitive validity of the scale may be called into question (Burton-Jones, 2009; Campbell & Fiske, 1959; Doty & Glick, 1998; Podsakoff et al., 2003). But what constitutes adequate degrees of cognitive validity? And how can cognitive validity be consistently and reliably assessed?

While scale adaptations appear to be frequent in practice, there has been no discussion: of the consequences of scale adaptation on the cognitive validity of IS measures, and of whether the methodological toolkit currently available to support scale adaptation practices is adequate. Meanwhile, there are many reasons why IS researchers will want to continue adapting existing scales for use in their research. For example, the ever-changing nature of IS-related phenomena demands that item wording is updated to reflect characteristics of contemporary technology affordances and organisational environments (Compeau et al., 2022). Another reason is that items demand to be routinely modified to reflect the specificities of the situational context of research studies. Therefore, while scale adaptation is not necessarily a bad practice, any adaptation necessarily raises questions about whether the adapted scale still does a good job of measuring the intended construct.

This paper adopts an assumption-challenging approach towards scale adaptation practices (Alvesson & Sandberg, 2011; Chatterjee & Davison, 2021). To question the assumptions that underly scale adaptation is relevant because this activity has settled as a seemingly opaque and inconsistent practice despite being frequently undertaken in IS. In this introduction, we identify five key beliefs underlying scale adaptation activities that we consider to be fallacious. We then provide evidence for why these beliefs ought to be challenged and outline the directions that this paper offers in response to these challenges. We summarise these points in Appendix A.

First, scale adaptation may be considered a trivial practice that involves only tangential item wording modifications (e.g., to modify the name of the target IT). While some adaptations may fall into this category, we find that significant departure from the baseline scale are quite common. In fact, our review of IS measure development practices (p. 6–10) shows that a third of adapted scales demonstrate less than 50% semantic similarity with the baseline scales from which they are derived, suggesting that crucial item wording modification are undertaken. Substantial modifications of items not only demand to verify that the scale still measures what it purports to measure; it also requires verifying that the modifications did not lead to the accidental introduction of PICs. This observation

motivates the development of a method that researchers could use to gauge the performance of an adapted scale from a cognitive validity standpoint.

Second, items that prospective respondents can smoothly process are often assumed to be free of PICs. We advance that this fallacious belief reflects a partial understanding of the cognitive validity concept. Indeed, there is a consensus in the literature that response effects can be observed even when respondents effortlessly process item content (Podsakoff et al., 2003; Tourangeau et al., 2000). For example, replacing a neutral term with an emotionally charged term in an item would not undermine its readability, but it could systematically influence responses due to the demand effects generated by the loaded word (Viswanathan, 2005). In response to this challenge, we provide a comprehensive definition of the cognitive validity concept, outline the mechanisms through which it can impact empirical results, and differentiate cognitive validity from the related concept of content validity.

Third, one may argue that current practices are effective in dealing against the proliferation of poorly adapted scales, and that the status quo situation is satisfactory. If this were the case, eminent IS scholars would probably be less vocal about the need to refocus on the fundamentals of measurement practices. Less than a decade ago, Gregor and Klein (2014) lamented about the proliferation of poorly worded items in IS research and on their detrimental impact on the theory-testing genre. Burton-Jones and Lee (2017) also observed that "researchers do not have good definitions of measures and measurement, nor do they have a clear agreement about how best to engage in or assess these activities" (p. 465), indicating that the issue has yet to be overcome. In fact, our review of measure development practices in IS reveals that cognitive validity is formally tested in less than 2% of the papers that use self-reported measures (cf. Table 1). This parallels similar findings that measure development practices are "inconsistent, fragmented, and incomplete" (Schmitz & Storey, 2020, p. 43). Undeniably, there are some ambiguities in the way cognitive validity should be assessed, and in how the empirical data resulting from this activity should be interpreted. We believe that making the execution and interpretation of cognitive validation more transparent, consistent, and reliable could contribute to more frequent assessments.

This leads us to the fourth belief that this paper wishes to challenge: the idea that problematic items resulting from scale adaptations are easy to detect and that PICs can effortlessly be diagnosed using existing techniques (i.e., expert reviews, field pretests, focus groups, interviews). Contrary to this belief, the literature indicates that detecting relevant wording problems is a challenging task, because even subtle alterations can influence how people process item content and form a response (Bradburn et al., 2004; Dillman et al., 2014). Consequently, although a great number of problems may be unearthed by experts, few of them actually matter in practice (Diamantopoulos et al., 1994). Empirical observations of item review panels confirmed that experts seldom converge on their conclusion regarding problematic items, suggesting that detecting problematic items is far from a straightforward task (cf. Study 2). To increase the reliability of problematic item detection activities, this paper develops a method that minimises the influence of subjective judgements at both the data collection (large pool of judges) and statistical analysis (use of ANOVA) stages.

Last but not least, a taken for granted assumption underlying scale adaptation practices is that the impact of poor adaptations is confined to the study adapting the scale. Arguably, this belief may not hold in a discipline that encourages authors to expand previous research in a cumulative tradition (Keen, 1980). By routinely relying on each other's measures to capture latent constructs, PICs that are introduced at one stage of an adaptation may diffuse across multiple studies as a result of subsequent adaptations. Heggestad et al. (2019) refer to this phenomenon as "cascading adaptations", where "an author adapts a scale, and then subsequent authors use the adapted scale, citing the work of the author who adapted it" (p. 2614). Although we currently have a limited understanding of the mechanisms through which PICs may diffuse in the IS research field, it is reasonable to advance that adapted scales are reused by others. A new statistic for benchmarking the cognitive validity of adapted scales against venerable IS measures is introduced to curb the diffusion of problematic adaptations.

In summary, this paper makes two contributions to the IS literature. First, it advances our theoretical understanding of the cognitive validity concept in the context of frequent adaptations of IS scales. Second, it expands the methodological toolkit with a new method that can supplement authors' measure development efforts. Through these two contributions, this paper adds to the debate on measure development practices in IS research (Boudreau et al., 2001; Burton-Jones & Lee, 2017; Schmitz & Storey, 2020; Straub, 1989). It also more specifically answers recent calls for more explicit scale adaptation standards (Compeau et al., 2022; Heggestad et al., 2019).

The paper is structured as follows. First, we review measure development practices in IS to examine the extent of scale adaptations and to track how the IS field has dealt with the issue over the past decade ($n = 322$ articles). We then revisit scale adaptation practices through the lens of Tourangeau et al.'s (2000) model of survey responding, provide an operational definition of cognitive validity, and explain the mechanisms through which it impacts empirical results. The remainder of the article concentrates on the development and validation of a cognitive validity assessment method. Through two experiments, we show evidence of the ability of the method to detect PICs from subtle variations in item wording (Study 1) and we demonstrate that the new method can supplement item review panel exercises (Study 2). Next, we subject 10 venerable IS scales to the method and generate a new index to benchmark the cognitive validity of adapted scales across studies (Study 3). Altogether, these three studies provide converging evidence that the method can benefit the IS field by making the scale adaptation process more robust, transparent, and consistent. Building on this empirical work, we discuss the implications of the new method for IS research practices and provide directions for future research on measurement in IS.

2 | BACKGROUND AND MOTIVATION

2.1 | Scale adaptation in IS research

This paper focuses on scale adaptation that involves purposeful alteration of items' linguistic properties, which linguists break down into three different aspects (Clifton & Duffy, 2001): items' *lexical* structure (the words being employed in an item), *syntactic* structure (the arrangements of words and their relationships within an item), and *semantic* structure (the semantic space that an item is projecting).

There are reasons to believe that items' linguistic is routinely modified in IS research. First, given the relevance of some of the reference work to IS scholarship (Baskerville & Myers, 2002; Galliers, 2003), measures are often borrowed from domains such as applied psychology, marketing, accounting, communication, or computer science, and imported into the IS domain. To transfer a measure from another domain into IS usually requires altering its content so that it reflects IT-specific traits or behaviours. Second, IS phenomena are typically fast evolving, and the premises on which measures were created may become increasingly irrelevant over time (Compeau et al., 2022). As items may reflect outdated assumptions about the technology, people, or organisations, IS scholars will be pressed by the research field to revise them. Third, existing IS scales may be adapted to enhance the fit between the research instrument and the situational context of a particular study. Arguably, this type of adaptation is on the rise due to IS measures being developed and deployed across an ever-increasing range of contexts and spheres of life (Rai, 2020; Yoo, 2010).

While it seems reasonable to expect that alterations of item wording are commonplace in IS, our understanding of this practice remains limited with regards to its frequency and extent. Besides, it is unclear how IS scholars ascertain the cognitive validity of adapted scales. A review of current practices in IS research is conducted in the next section to shed light on these issues.

2.2 | Review of current practices in IS research

2.2.1 | Frequency and extent of scale adaptations

This review covered eight premier IS journals,[1] the years 2010 to 2020, and was performed on multiple databases to increase coverage.[2] Keywords included the terms "instrument," "construct," "scale," "measurement," and "measure". It was supplemented by a quick scan of the article to ascertain that the study employs self-reported measures. This search yielded a sample of 322 IS papers. In the following step, we examined the measure development sections of these papers. A preliminary examination revealed that 85% of the papers in our sample use a mix of existing scales, adapted scales and newly developed scales. Only 5% solely employ existing scales, and 10% rely solely on self-developed scales. This observation confirms that scale adaptation is a frequent practice in the IS field.

We also observed that scale adaptation (or sentences such as "measure XYZ was adapted from ABC") was used as an umbrella term reflecting very different realities. To further advance our understanding of this phenomenon and its nuances, we analysed the measure development section of these articles. Our objective was to assess the extent to which the adapted scales departed from the original baseline scales. For this task, we focused on the 18 papers in our sample that have been published in 2020, the most recent year in our search time frame. Whenever an already published scale had been adapted, we located the baseline scale from which the items originated, and we recorded the items from both the baseline and the adapted scale. Consistent with our definition of scale adaptation, adapted scales that did not modify the wording of the baseline scale were not retained.[3] A total of 223 items from 54 scales were recorded using this process.

To assess the extent to which an adapted scale departed from its baseline, we subjected both versions of the scale to the http://wordvec.colorado.edu website, which provides an automated latent semantic analysis (LSA) tool to assess the extent of semantic similarity between two documents. LSA is a natural language processing (NLP) technique that compares two vectors in a high-dimensional space, with the cosine of the angle between the two vectors representing the extent of the semantic similarity of the two texts (Gefen et al., 2017; Gefen & Larsen, 2017; Landauer et al., 2013). As a cosine measure, its absolute value ranges between 0% and 100%, with 100% representing complete similarity, and 0% representing no similarity. Higher cosine scores between text segments signify higher semantic overlap. We performed pairwise comparisons of the baseline scales against their adapted scales using the LSA tool. For quality insurance purposes, we also compared the semantic similarity of the baseline scale with an unrelated "Belief in Fate" (BIF) scale (Chan et al., 2009). Given that the BIF scale does not share any conceptual or empirical meaning with the shortlisted IS scales, we expected similarity scores involving the BIF scale to approximate 0%. The results of this comparison are reported in Appendix B.

The similarity scores of the baseline scales against the adapted scales reveal moderate degrees of semantic overlap with an average similarity of 62% (min = 13%; max = 92%; SD = 22%). In contrast, the average degree of similarity between baseline scales and the BIF amounted to only 9% (min = 1%; max = 32%; SD = 8%), which is indicative of the validity of the test. Eight adapted scales demonstrate high degrees of similarities with their baseline (similarity >80%). Scales with high degrees of similarity tend to be those that make a minor adjustment to item wording. For example, the "Perceived increase in job complexity" scale (Zimmermann et al., 2020) which has received the highest

[1]Information Systems Journal (ISJ), European Journal of Information Systems (EJIS), Information Systems Research (ISR), Journal of the Association for Information Systems (JAIS), Journal of Information Technology (JIT), Journal of Management Information Systems (JMIS), Journal of Strategic Information Systems (JSIS), and MIS Quarterly (MISQ).

[2]Springer, EBSCO, Sage, Science Direct, JSTOR, Informs, Google Scholars, Taylor & Francis, AIS eLibrary, and Wiley.

[3]It is important to note this data collection remains partial, as our criteria for reporting a scale could not accommodate every adaption we encountered. Our criteria were as follows: (1) a maximum of three adapted scales per paper were retained to ensure a balanced sample; (2) whenever more than one scale was cited as a baseline, we recorded the items from only one baseline scale; (3) when three or more baseline scales were cited as sources, we did not record the adaptation, because we considered this practice more akin to construct mixology than to scale adaptation (Newman et al., 2016); (4) only Likert-type scales were retained (several semantic differential scales were not recorded).

similarity score in our sample (92%) has substituted the item "As a result of offshoring the job now requires that I only do one task or activity at a time" to the item "The job requires that I only do one task or activity at a time". In contrast, some of the scales in our sample departed significantly from the baseline. This is the case of the "Inter-firm Knowledge Distance" scale (Cui et al., 2020), the "Security response efficacy" scale (Silic & Lowry, 2020), or the "Online Post-Transaction Effort Convenience" scale (Trenz et al., 2020) (13%, 30%, and 30% similarity, respectively). These results show that adapted scales may depart from their baseline in a critical way, leaving room for the introduction of PICs.

2.2.2 | Activities undertaken to examine the cognitive validity of IS measures

A follow-up question pertains to the methods that are executed by IS researchers to minimise the presence of PICs in adapted scales. To be more conclusive on this aspect, we turned back to our sample of 322 papers, and systematically coded (1) whether the paper explicitly discussed the cognitive validity issue, (2) whether it took steps to attend to problematic item wording, and (3) what method(s) had been implemented to deal with the issue, if any.

We found that 34% of the papers explicitly refer to at least one cognitive validity issue. Usually, these discussions pertain to problems of double-barreledness, vagueness, or ambiguity, that arise at the comprehension stage. Nearly all of the papers discussing a cognitive validity problem (32%) implemented at least one method to examine the cognitive validity of their scale. The most common methods implemented to examine cognitive validity are expert reviews (15%), pilot studies (12%), field pretests (10%), card sortings (4%), back-translations (4%), and interviews (3%). More targeted procedures such as cognitive interviewing and item clarity ratings are seldom used (2% and close to 0%, respectively). 15% of the papers used only one of the above-mentioned techniques, whereas 17% combined two or more of them. These results are summarised in Table 1.

This review of measure development practices is informative with respect to two considerations. First, it provides evidence that substantial modifications of item wording are frequent, and some are akin in magnitude to creating new measures. This was reflected in the moderate similarity scores found between the adapted scales and baseline scales. However, this finding must be nuanced, as several adaptations entailed only marginal wording modifications, reflecting important variability within and across studies when it comes to adapting existing scales. Second, this review shows that although approximately one-third of the IS papers in our sample proactively take steps to improve the cognitive validity of their measures in one way or another, these efforts overwhelmingly concentrate on item comprehension issues. Besides, very few papers have implemented formal analytical tests of cognitive validity,

TABLE 1 Activities undertaken to examine the cognitive validity of IS measures

	Frequency	%
Uses self-reported measure (Total)	322	100
Mentions the issue	110	34
Takes formal steps	104	32
Translation	12	4
Expert review	48	15
Interviews	9	3
Card sorting	13	4
Field pretest	33	10
Pilot study	40	12
Cognitive interviewing	5	2
Item clarity rating	1	0

such as item clarity ratings or cognitive interviewing, corroborating what other recent reviews have observed (Schmitz & Storey, 2020). One notable exception was found in (Sun, 2012), where the author asked a pool of 15 subjects to rate the clarity of each of their scale items (1 for clear and 0 for unclear).

These observations parallel Burton-Jones and Lee's (2017) suggestion that IS researchers tend to overlook the measure development phase because they lack awareness of the "full landscape of measurement issues" (p. 464) and lose sight of the most fundamental problems with measurement. This confirms our observation that efforts to assess cognitive validity are generally opaque and inconsistent. In addition to the ambiguities that pertain to the assessment of cognitive validity, there are some lingering questions concerning the reliability of the available methods. For example, expert reviews and focus groups may be subject to group think or expert biases (Schriesheim et al., 1993). Similarly, the literature suggests that independent researchers would likely reach different conclusions when subjecting the same scale to the cognitive interviewing procedure (DeMaio & Landreth, 2004; Forsyth et al., 2004; Rothgeb et al., 2007). All these issues pertain to the cognitive mechanisms that underpin the survey response process.

2.3 | Examining scale adaptation practices through the lens of the survey response process model

Tourangeau et al. (2000) model of survey response is the most comprehensive and consensual model to date to appreciate how alterations of items' linguistic structure can influence how respondents process survey items. This model identifies the four mental stages through which respondents go as they process survey items: (1) *comprehension* (i.e., determining what the question is asking), (2) *retrieval* (i.e., recalling the relevant information from memory), (3) *judgement* (i.e., processing the information to formulate an answer), and (4) *response selection* (i.e., mapping that answer onto the provided response categories). As summarised in Table 2, each stage can be divided into specific mental subprocesses. Several PICs can interfere with these cognitive operations and undermine the integrity of the survey response process. We briefly detail these four stages, the interferences that PICs could create, and outline illustrative areas of IS research where we think these problems would be particularly salient.

2.3.1 | Comprehension stage

This stage is concerned with respondents working to understand the focus or intent of the item. This involves processing individual words and constructing a general understanding by identifying the syntactic structure that glues them together. Item ambiguity and item complexity can undermine respondents' understanding of what the item asks and prompt them to ascribe their idiosyncratic meaning to a scale's items. In strategic IS research dealing with abstract organisational concepts, the use of terms such as "flexibility" (Chengalur-smith et al., 2010), "knowledge" (Wang et al., 2008) or "capabilities" (Chen et al., 2010) can obfuscate the meaning of the items and confuse respondents who could then provide a weakly grounded answer.

2.3.2 | Retrieval stage

Once respondents have a good grasp of what the item is asking, they will gather relevant information from memory to formulate an answer. The wording of the item will heavily influence what strategies respondents employ to retrieve information, and whether they will search for generic or specific memories. Problems may arise at this stage with items that make erroneous assumptions about what respondents know, or when they incite them to retrieve

TABLE 2 Survey response process model and IS examples

Stage of the response process	Operations performed at each stage	Problematic item characteristics (PICs)	Examples in IS research
Comprehension: the respondent works to understand what is asked	• Determine what the item is asking • Decide what individual words mean • Figure out what the survey is asking for	• Item ambiguity (i.e., vague semantic meaning, double-barreled statement, equivocal meaning) • Item complexity (i.e., item length, complex syntactic structure)	*Strategic management of IS*: respondents may be unsure about the meaning of abstract concepts such as "strategic capabilities" or "increased flexibility" that are used in strategic IS research instruments
Retrieval: the respondent gathers relevant information	• Work general strategies and cues • Retrieve specific and generic memories • Fill in missing details	• Item contains erroneous assumptions (i.e., cognitive ability, knowledge on the topic) • Item demand characteristics (i.e., leading statement, implied answer)	*IT usability*: respondents' ability to answer IT usability questions is a function of the recency and depth of their behavioural interaction with technology, which researchers could overestimate
Judgement: the respondent formulates an answer	• Determine completeness and accuracy of retrieved memories • Integrate the retrieved material or make inferences that fill in the gaps in retrieval	• Item demand characteristics (i.e., leading statement, implied answer) • Item social desirability (i.e., sensitive questions, moral/ethical issues, socially valued attributes)	*Dark side of IT use*: respondents may overstate the negative consequences of technology use in response to emotionally charged and negatively valenced items
Response: the respondent reports his or her answer	• Convert answer into required format • Map judgement onto response category • Edit response for consistency, acceptability, or other criteria	• Item demand characteristics (i.e., leading statement, implied answer) • Item social desirability (i.e., sensitive questions, moral/ethical issues, socially valued attributes)	*Information security*: respondents may underreport behaviours that could have resulted in leakages of copyrighted material or other critical business information

Source: Adapted from Dillman et al. (2014), Tourangeau et al. (2000), Podsakoff et al. (2003).

certain specific memories that result in partial or biased information when memories are pieced together. This problem could arise in IT usability research for example, when the ability of respondents to recall information is directly related to the recency and depth of their behavioural interaction with the target technology (Hoehle & Venkatesh, 2015; Winter et al., 2003). Researchers may overestimate the recency and accuracy of respondents' memory by assuming that the target system of the questionnaire is being used on a regular basis.

2.3.3 | Judgement stage

Although some items may require reporting estimates of behavioural or observable information (e.g., performance figures), most of the time respondents are asked to report their opinion or belief on a given issue. Thus, once respondents have searched memory for information relevant to the item, they will integrate the retrieved material and start forming an overall judgement. In case important information is missing, they will make inferences aimed at filling possible gaps in memory. Demand effects embedded in item wording as well items that call for socially desirable

answers can influence the judgement formation process and result in distorted perceptions of the reality of a given situation. The stream of IS research dealing with negative and unexpected outcomes of IS design, implementation, and use is particularly subject to this issue. In this context, respondents could be inclined to overestimate the negative consequences of IT in response to emotionally charged terms, or under-report certain behaviours if they believe that those are negatively perceived by their peers.

2.3.4 | Response stage

This stage is concerned with the respondents deciding which response option is best suited to collect their answer. With Likert scales, options usually vary from "*strongly disagree*" to "*strongly agree*", but alternative formats such as semantic differential scales may also be employed to capture respondents' answers (Chin et al., 2008). Conscious or unconscious distortions may occur at this stage, leading respondents to report a score that is not aligned with the judgement that they have formed previously. Item demand effects and socially desirable wordings can prompt respondents to select an answer that differs from the one they initially intended to. These effects are reportedly more pronounced in research on IT use and addiction, or in information privacy and security research (Kwak et al., 2019). In their study on cyberslacking in the workplace, which requires respondents to report the frequency at which they may engage in deviant behaviours that violate the social norm, Venkatesh et al. (2023) acknowledge that "as this work deals with negative workplace behavior, the participants may not have answered questions truthfully due to social desirability." (p. 40).

2.4 | An operational definition of cognitive validity

To build an operational definition of cognitive validity, we reviewed the survey measurement literature to determine what PICs may be detrimental to cognitive validity. In this review, we focused on item *characteristics*, namely higher-level aspects that may be subjectively associated with items (e.g., ambiguous, complex, sensitive, leading, etc.), rather than item *features* that tend to be observable elements of items (e.g., item length, presence of negations, indeterminate frequency qualifiers, etc.). Our review focused on the literature on the cognitive aspects of survey methods (CASM) that has emerged in the 80's with the objective of illuminating the cognitive and communicative processes underlying survey responding (Bradburn et al., 2004; Krosnick, 2018; Schwarz, 2007; Schwarz & Sudman, 1996; Sudman et al., 2010; Tourangeau, 2018; Vannette & Krosnick, 2018). We combined this literature with the more specific literature on measurement in management (Podsakoff et al., 2003) and IS research (Burton-Jones & Lee, 2017; Burton-Jones & Straub, 2004; Schmitz & Storey, 2020; Straub & Gefen, 2004).

Three broad categories of PICs emerged from this search: the fact that items may be written in a way that is difficult to understand intuitively, that items may be written in a way that reflects socially desirable aspects, and that items may contain cues that reveal how to respond to them. In addition to these categories, the survey methodology literature suggests that researchers and investigators often make erroneous assumptions about what respondents know, forcing them to take a wild guess when providing an answer. This observation, which resonates with the concept of "knowledge bias" identified by Burton-Jones (2009), led us to consider a fourth type of PIC to capture the fact that items may be worded in a way that requires thorough knowledge of the IS domain to produce an appropriate response.

Consequently, this review yielded a framework structured around four main types of PICs: comprehension difficulty (CD); social desirability (SD); response inducement (RI); and knowledge assumption (KA). These four categories were further broken down to capture important nuances that emerged from the literature review. The PICs that can potentially undermine the cognitive validity of IS measures are summarised in Table 3 and detailed below.

TABLE 3 Problematic item characteristics (PICs) altering the cognitive validity of measures

Category	Definition	Key references
Comprehension Difficulty: *refers to the fact that the focus of items may be difficult to intuitively grasp*		
Ambiguity	The extent to which item content is vague or subject to multiple interpretations	(Fowler, 1992; Hardy & Ford, 2014; Johnson, 2004)
Complexity	The extent to which item content is cumbersome for respondents to process	
Social Desirability: *refers to the fact that items may be written in such a way that reflects more socially desirable attitudes, behaviours, or perceptions*		
Moralistic bias	The extent to which item content induces respondent to construct "saint-like" images of themselves	(Kuncel & Tellegen, 2009; Nederhof, 1985; Paulhus, 2001; Steenkamp et al., 2010)
Egoistic bias	The extent to which item content induces respondents to exaggerate their social and/or intellectual status	
Response Inducement: *refers to the fact that items may convey hidden cues for how to respond to them*		
Suggested response	The extent to which item content reveals the personal preference of its author/the researcher	(Johnson & Christensen, 2008; Knowles & Condon, 1999; Stalnaker, 1977)
Implied response	The extent to which item content logically implies a particular response or a response direction	
Knowledge Assumption: *refers to the fact that items may require thorough domain knowledge to produce adequate responses*		
Practical	The extent to which item content suggests that respondents have thorough practical knowledge in the IS domain	(Blair & Burton, 1987; Burton-Jones, 2009; Wilson & Dunn, 2004)
Theoretical	The extent to which item content suggests that respondents have thorough theoretical knowledge in the IS domain	

2.4.1 | Comprehension difficulty (CD)

CD refers to the aspects that undermine the straightforward understanding of item statements by the respondents. CD manifests in items that are characterised as either *Ambiguous* or *Complex*. Item ambiguity is the "uncertainty about the denotative reference or meaning of an item" (Johnson, 1986), suggesting a problem of a semantic nature. Indeed, item ambiguity may have its roots in the vagueness of an item's denotative meaning (Hardy & Ford, 2014), the existence of multiple conflicting meanings (Fowler, 1992), or the abstractness of the underlying construct (Doty & Glick, 1998), all of which relate to underlying semantic problems. In contrast, the issue of item complexity has a syntactic or grammatical nature. That is, although respondents may be unable initially to efficiently process the item content, they may understand its intended meaning after multiple successive iterations. The source of item complexity may be inadequate grammar or complicated wording caused by double-barreled assertions, a wordy formulation, or the use of negative assertions that exacerbate the complexity of items.

2.4.2 | Social desirability (SD)

SD refers to the fact that items may be written in such a way that they reflect socially desirable attitudes, behaviours, or perceptions (Kuncel & Tellegen, 2009). Socially desirable items tend to lead to socially desirable responses,

whereby respondents attenuate or accentuate their reporting of certain traits, attitudes, or behaviours (Nederhof, 1985). Paulhus (2001) provides the most comprehensive model to date for understanding the tenets of socially desirable responding. He identified two primary forms of socially desirable responding: *Moralistic bias* and *Egoistic bias*. Moralistic bias is an exaggerated sense of moral standing. Deception may be either unconscious, which leads to the denial of one's faults ("self-deceptive denial"), or more deliberate, which minimises one's mistakes or harm towards others to maintain a positive image ("communion management"). Egoistic bias refers to having an inflated opinion of one's social or intellectual status. As with moralistic bias, this tendency to claim overly positive characteristics of oneself may be sincere and unconscious ("self-deceptive enhancement") or more instrumental and strategic through bragging or self-promotion ("agency management"). Although socially desirable responding correlates with personality attributes and ultimately originates from the respondent (Crowne & Marlowe, 1960), researchers should strive to craft items that attenuate the presence of socially desirable cues that stimulate socially desirable responses (Kuncel & Tellegen, 2009).

2.4.3 | Response inducement (RI)

RI refers to the fact that items may convey subtle cues as to how to respond to them. Although this category commonly refers to leading questions, one may conceive of different RI levels according to the intensity of the demand that they induce. Thus, we distinguish between item statements that *suggest* responses (low-intensity demand) and statements that logically *imply* responses (high-intensity demand). Items that *suggest* responses subtly lead respondents to consider a set of responses superior to other responses by attracting the attention of the respondent. Such items contain features that tend to betray the personal opinion or preference of the person who developed them (Stalnaker, 1977). In the case of agreement scales, items that suggest a response tend to encourage respondents to uncritically endorse a statement (Knowles & Condon, 1999). Similarly, items that contain emotionally charged terms create strong affective reactions that lead respondents to consider specific options (B. Johnson & Christensen, 2008). In contrast, implied responses gather item statements that logically imply a specific response or a more general response direction, which occurs when selecting a response direction that differs from the one implied by the item would be a blatant logical mistake. In both cases, the consequence of RI is an artificial narrowing of the range of allowable responses.

2.4.4 | Knowledge assumption (KA)

KA emerges when items require thorough domain knowledge to produce adequate responses. This assumption manifests when people are asked to report behaviours that they have forgotten about (Blair & Burton, 1987) or when they are not aware of their own mental states (Wilson & Dunn, 2004). When the target of the questionnaire is extrinsic to respondents, researchers may assume that respondents know more than what they report (Payne, 1951). In IS, researchers typically assume that respondents have sufficient knowledge about the tasks, processes, structures, and/or technology, depending on the focus of the research. However, these assumptions could well exceed what respondents actually know, thus altering the validity of the measurement (Burton-Jones, 2009). These assumptions include overestimating the respondents' general degree of competency with a technology, the respondents' degree of embeddedness within an IS, or the depth with which respondents interact with the system (de Reuver & Bouwman, 2015). For example, Collopy (1996) found that computer use is not well estimated by users because infrequent users overestimate and frequent users underestimate usage levels. To refine this category, we introduce an analytical distinction between IS domain knowledge gained through *practical* experience (e.g., direct behavioural interactions, training sessions, hands-on expertise) and knowledge gained indirectly through *theoretical* means (e.g., press magazines, word-of-mouth, self-research). This distinction draws on the conceptual difference between

"learning by doing", which occurs by first-hand or direct experience with technology, and "learning from others", which occurs by vicarious or indirect encounters, such as verbal descriptions, depictions, or modelling (Ryu et al., 2005).

2.5 | Impact of cognitive validity in IS survey research

The operational definition of cognitive validity allows us to anticipate on the impact that inadequate attention to cognitive validity could have on a field that is compelled to continuously revisit the wording of its measures. Four primary areas of impact are considered: content validity, reliability, validity, and common method variance. These areas are frequently cited as important considerations to pay attention to in survey research when establishing the validity of empirical work (Bagozzi, 2011; MacKenzie et al., 2011). We review these four areas and briefly discuss the mechanisms through which cognitive validity can impact them. The essence of this discussion is summarised in Table 4.

First, inadequate attention to cognitive validity could be detrimental to the content validity of IS measures, which refers to the extent to which a researcher's conceptualization of a construct is reflected in its operationalization of it (Moore & Benbasat, 1991; Schmitz & Storey, 2020). There is an intimate connection between cognitive validity and content validity because both issues are rooted in items' linguistic characteristics. When assessing content validity, researchers' key concern is on whether a scale's items convey the meaning of the intended construct over and above that of related constructs (Colquitt et al., 2019). If a scale's items are worded in a complicated manner (CD), the intended meaning of the items will likely be obfuscated, which would undermine the ability of the scale to precisely reflect the content domain of the underlying construct. On the other hand, scales that induce socially desirable responses (SD) will likely capture aspects that are not part of the construct (such as respondents' tendency to overstate things) (Steenkamp et al., 2010). Similarly, when the wording of the items is leading (RI), the scale will capture respondents' tendency to acquiesce with the keying of the scale (Baumgartner & Steenkamp, 2001). Insufficient content validity has been linked to measurement deficiency (i.e., the measure does not capture all aspects of the construct) and measurement contamination (i.e., the measure captures aspects that are not part of the construct) (Schmitz & Storey, 2020).

PICs also represent a liability to the reliability (i.e., consistent measurement of the intended construct) and validity (i.e., accurate measurement of the intended construct) of IS measures (Campbell & Fiske, 1959; Nunnally & Bernstein, 1994). Indeed, PICs can produce response effects that translate into either random or systematic measurement error (Viswanathan, 2005). Given that reliability is concerned with the extent of random error contained in a measure, and that validity pertains to the extent of both random and systematic error contained in a measure (Viswanathan, 2005), both validity and reliability are likely to be impacted by PICs. Specifically, random error would accrue from ambiguous items that respondents fail to correctly understand (CD) or from items that force respondents to take a wild guess because it is assumed that they know more than they actually do (KA). Meanwhile, systematic error would be caused by respondents providing inaccurate but consistent responses, as is the case with items plagued with RI or SD. Increased amount of random error could lead to poor internal consistency reliability (e.g., Cronbach alpha) or the scale not performing consistently over multiple administrations (test–retest reliability). On the other hand, increased amounts of random or systematic error would impact factor loadings and dimensionality (EFA stage), or overall fit indices, loadings, and residuals (CFA stage).

A fourth foreseeable implication of inadequate attention to cognitive validity lies in the extent of artifactual covariance contained in measurement and structural models. In their seminal work on common method effects, Podsakoff et al. (2003) have warned against the salience of "item characteristic effects" in management research, namely "any artifactual covariance that is caused by the influence or interpretation that a respondent might ascribe to an item solely because of specific properties or characteristics the item possesses" (p. 882). Item characteristics effects alter (usually inflate) both inter-item and inter-construct correlations and could lead researchers to erroneously conclude that a research model is supported (Sharma et al., 2009). While method variance is commonly viewed from the broader perspective of the method, our investigation invites us to think of the issue from the micro vantage point of the individual

TABLE 4 Implications of inadequate attention to cognitive validity on empirical IS research

Areas of impact	Facets of cognitive validity				Implications for empirical IS research
	CD	RI	SD	KA	
Response stage	Comprehension	x			
	Retrieval		x		x
	Judgement		x	X	
	Selection		x	X	
Content validity	Obfuscates the meaning of the items	Increases likelihood of capturing unrelated aspects (e.g., acquiescent response style)	Increases likelihood of capturing unrelated aspects (e.g., overclaiming tendency)	N.A.	Measurement deficiency/ contamination
Reliability	Increases random error	N.A.	N.A.	Increases random error	Poor internal consistency reliability, insufficient test–retest reliability
Validity (*)	Increases random error	Increases systematic error	Increases systematic error	Increases random error	EFA: weaker loadings, dimensionality issues CFA: increased residuals, decreased model fit
Common method effects	Inflates inter-item and inter-construct correlations	Inflates inter-item and inter-construct correlations	Inflates inter-item and inter-construct correlations	Inflates inter-item and inter-construct correlations	Inflated measurement or structural model properties

Note: (*) Convergent validity, Discriminant validity, Nomological validity, Predictive validity.
Abbreviations: CD, comprehension difficulty; CFA, confirmatory factor analysis; EFA, exploratory factor analysis; KA, knowledge assumption; RI, response inducement; SD, social desirability.

measure (Spector et al., 2019). Therefore, when all of the items of the same scale share a cognitive validity issue, they share a common method factor that will inflate inter-item correlations. Similarly, if the same cognitive validity issue is shared across several constructs, relationships among those constructs will be artificially inflated.

The cognitive validity definition along with the assessment of its impact on empirical IS research synthesise and integrate key problems pertaining to item wording effects. However, realising the full potential of these developments would require creating a methodological process that can inform scale adaptation practices. This leads us to the second objective of this paper, which is to propose a new method to supplement measure development processes.

3 | METHOD DEVELOPMENT

We build on the conceptual framework introduced in Table 3 to develop a method to assess the cognitive validity of IS measures. The method hinges on the rating items by naïve judges on the four primary categories and eight subcategories defined in the previous section. The method development process followed a four-step approach: (1) we operationalised a data collection protocol and pretested it to verify the feasibility of the new method; (2) we tested the validity of the method using one-way repeated ANOVAs with planned contrasts (Study 1, $N = 34$ judges); (3) we verified the utility of the method through an observation study involving four independent expert panels (Study 2, $N = 17$ researchers); and (4) we subjected 10 venerable IS scales to the method to clarify the criteria for interpreting the resulting cognitive validity statistics (Study 3, $N = 479$ judges). A breakdown of the objectives and activities of this research program is provided in Table 5 and subsequently detailed.

3.1 | Step 1-operationalization of the cognitive evaluation procedure

3.1.1 | Developing the data collection procedure

We started by creating a data collection protocol geared towards the detection of problems in item wording. Key requirements for the protocol were as follows: (1) judges without specific training in questionnaire design and psychometric measurement should be able to confidently perform the task, (2) transparency in the data collection protocol should allow for the production of a clear audit trail, and (3) numerical scores that could be analysed using standard statistical techniques could be produced (e.g., *t*-test or ANOVA). With these requirements in mind, we developed the core features of the data collection protocol, starting with the generation of a set of measures to gauge the salience of PICs in items (i.e., PIC measures). The PIC measures were deductively derived from the operational definition of cognitive validity (see Table 4). At this stage of the method development process, the PIC measures were phrased very differently and used alternative response anchors. Instructions to inform the judges about the nature of their assessment were also drafted.

3.1.2 | Pretesting the data collection procedure

Once developed, the draft procedure was pretested with 4 Ph.D. candidates to verify the integrity of the data generation protocol. After a brief introduction of the purpose of the new method, the instruction was given to the participants to assess the items and report their rating using the provided assessment sheet. Participants were encouraged to raise their questions as they went through the procedure, and clarification was provided when needed. Once the task was completed, a collective debriefing was organised to discuss the problems that surfaced. The entire pretest was fully video recorded and further discussed within the research team. Modifications targeted at minimising the effort required to perform the rating task and improving the reliability of the procedure were implemented.

TABLE 5 Overview of the method development process

Steps	Objectives	Activities
Step 1: Operationalization of the Cognitive Evaluation Procedure	Operationalise the assessment framework and pretest it to make the necessary adjustments	• Designed the data collection protocol and created the associated material • Pretested the data collection protocol with 4 Ph.D. students using a think-aloud approach • Performed a series of pilot tests in realistic data collection conditions
Step 2: Validation of the Cognitive Evaluation Procedure (Study 1)	Find evidence that the method can detect PICs	• Manipulated a series of items to introduce a range of PICs • Asked raters to assess cognitive validity using the pretested procedure • Tested differences in mean ratings using repeated ANOVAs with contrasts ($N = 34$ judges)
Step 3: Evidence of the Utility of the Procedure (Study 2)	Observe how researchers appropriate the method and derive value from it	• Designed a study protocol inspired by experimental methods and developed the study material • Subjected two scales to the previously validated cognitive validation procedure • Observed how four independent panels ($N = 17$ academic experts) appropriated the method in practice
Step 4: Evaluating Cognitive Validity in IS (Study 3)	Clarify the criteria for interpreting the numerical scores produced by the method	• Created a summary statistic to facilitate the interpretation of the raw ratings produced by the judges • Selected 10 venerable IS scales and subjected them to the method ($N = 479$ judges) • Produced a set of empirically derived cutoff points to help interpret the cognitive validity of IS measures

With a basic version of the rating procedure in place, the next step was specifically focused on (1) experimenting with alternative response selection formats and (2) verifying that the procedure could take place in realistic conditions, namely, without the guidance of a member of the research team. During the pilot phase, we also verified the ability of the method to accommodate a variety of IS construct and scale types (i.e., reflective and formative, macro construct and micro construct), the effectiveness of alternative rating formats (i.e., Likert, semantic differential), the wording of the PIC measures, and the clarity of the instructions.

We recruited judges using the online crowdsourcing platform Prolific (www.prolific.co). Because collecting rich feedback from the participants was deemed important at this stage, each rater was prompted to leave a written comment about the task that he or she undertook. In addition, the judges were asked to report how difficult the task was (*1-Extremely easy; 5-Extremely difficult*) and how confident they were in their assessments (*1-Not at all confident; 5-Extremely confident*).

3.1.3 | Results and discussion

The pretest and pilot phases yielded a mix of qualitative and quantitative feedback that was analysed to enhance the validity of the data collection procedure. One set of iterative improvements concentrated on the PIC measures. First, the PIC measures were aligned on a unique Likert-type response option format that minimised the cognitive burden

associated with shifting from one format to another. Second, the wording of the PIC measures was revised to reflect a general characteristic of the items rather than the presence of specific item features. For example, response inducement was initially detected using the measure "*The statement contains emotionally charged* terms" and was updated to the more encompassing measure, "*The statement contains clues that reveal the personal preference of its author*." Third, the PIC measures that required the judges to adopt the position of a prospective respondent were dropped. Indeed, measures such as "*Assuming that you have no experience at all with the target ICT, how confident would you be answering the statement?*" proved excessively taxing and confusing to the judges. More direct measures such as, "*The statement assumes that respondents have extensive hands-on IT experience*," were favoured.

Following the pretest and pilots, we noted a qualitative improvement in the judges' testimonies, some of them expressing unexpected levels of appreciation about the rating procedure once all of the changes were implemented.[4] Overall, this phase showed evidence for the feasibility of the data collection protocol and allowed us to calibrate the procedure appropriately.

3.2 | Step 2-validation of the cognitive evaluation procedure (study 1)

Step 1 was focused on the development of a procedure for evaluating the cognitive validity of measures. Step 2 is intended to find evidence that the method can detect the PICs present in items, thus providing evidence of the validity of the method.

3.2.1 | Research design

We expected evidence of the validity of the procedure to manifest in higher ratings for items that contain PICs over items that are free of PICs. To test this hypothesis, we employed a within-subject experimental design whereby each participant rated two versions of the same item: one version that was free of PICs (i.e., baseline item) and one version containing PICs (i.e., altered item). By exposing the judges to both the baseline and the altered version of an item, we can determine whether the judges would react differently to adapted items when they contain PICs.

Accordingly, we developed a list of item pairs, with each pair comprising a *baseline* version of the item and an *altered* version of the same item. To develop the item pairs, we first examined scales published in Basket of Eight IS journals and searched for items that mapped onto the assessment framework provided in Table 4. Then, we manipulated the wording of the items by accentuating (altered items) the salience of PICs. The manipulations were kept minimal to ensure that the meanings of the two items were nearly identical. Besides, to minimise the influence of word count as a confounding factor, the length of the altered items was kept as close as possible to that of the baseline items. Finally, two academic experts external to the team were informally consulted to verify the face validity of the manipulations, and minor changes were implemented based on their feedback. The item pairs and their sources are reported in Table 6.

3.2.2 | Data collection

We recruited US-based students aged between 18 and 25 years whose first language was English by using the crowdsourcing platform Prolific. We disseminated two instructional manipulation checks throughout the

[4]Selected testimonies: "Probably one of the more interesting surveys I've done, very nice!"; "[It] was an interesting task that differs from most surveys found on here"; "I enjoyed completing this task. It was simple to complete and easy to understand, especially with the explanations preceding the questions".

TABLE 6 Baseline and manipulated items

PIC category	Baseline item	Altered item (*)	Originating scale
Comprehension Difficulty (CD)			
Ambiguity	Our strategic decisions are made by people who belong to distinct departments	The level of participation in strategic decisions by diverse interests of the organisation is high	Planning Participation (Segars & Grover, 1999)
Complexity	Use of my current mobile phone has allowed me to try new work-related tasks.	I try hard to figure out how to perform work-related tasks in new places and settings that were not possible without my current mobile phone	Exploratory Task Adaptation (Schmitz et al., 2016)
Social Desirability (SD)			
Moralistic	I rarely contribute my knowledge to other people within my department	I *frequently* contribute my knowledge to other people within my department	Knowledge Sharing (Hsu & Chang, 2014)
Egoistic	I could hardly complete my job using this software even if someone showed me how to do it first	I could *easily* complete my job using this software even if there was no one around to tell me what to do	Personal Innovativeness (Agarwal & Prasad, 1998)
Response Inducement (RI)			
Suggested	I spend less time with my family because of this technology	I spend less *valuable* time with my family because of this technology	Techno-overload (Srivastava et al., 2015)
Implied	My unit frequently participates in customer-related activities (e.g., business and product planning)	My unit *actively* participates in activities that provide *excellent service* to our customers	Service Vision (Jia & Reich, 2013)
Knowledge Assumptions (KA)			
Practical	The software provides adequate guidance for me to navigate its content	The software provides *precise structural information* to locate content	Navigability (Jen-Hwa Hu et al., 2017)
Theoretical	I use my mobile phone in a manner that differs from the average user	I use my mobile phone in a manner that is *not encouraged by the designers*	Exploratory Task Adaptation (Schmitz et al., 2016)

Note: (*) Italics are used to emphasise the manipulated PIC.
Abbreviations: CD, comprehension difficulty; KA, knowledge assumption; RI, response inducement; SD, social desirability.

questionnaire (Oppenheimer et al., 2009) and automatically discarded the observations that failed them, which left us with a sample of 34 respondents. The individuals who had already participated in the pilot tests could not participate in the validation study. According to the central limit theorem, a sample size as small as 30 to 50 is considered adequate to assure a normal distribution of sample means in ANOVA, (Agresti, 2017; Vercruyssen & Hendrick, 2011). This is also the recommended sample size in similar data collection procedures that require the analysis of differences in item characteristics through comparisons of mean ratings (Hinkin & Tracey, 1999).

Each judge was asked to evaluate the cognitive properties of the baseline and manipulated items using the PIC measures pretested in Step 1 using a Likert-format scale with seven anchors (*1-Strongly disagree* to *7-Strongly agree*). Each item from the same pair was considered a distinct experimental condition to which the raters were exposed following a within-subject design. The test strategy then consisted of comparing the mean ratings of each *baseline* item to its *altered* counterpart. We considered the data collection procedure to be valid when the following two conditions were met: (1) the mean rating of the *baseline* item was lower than the mean rating of its corresponding *altered*

item, and (2) the difference in the mean rating between the *baseline* item and the *altered* item was statistically significant.

3.2.3 | Results and discussion

The results of Study 1 are presented in Table 7. A cursory review of the mean ratings indicates that the *baseline* items have systematically lower ratings than their *altered* counterparts, which satisfies the first criteria of validity. To test the significance of the differences in the mean ratings across the item pairs, we performed a series of repeated one-way ANOVAs with planned contrasts. Before computing the F-tests, we determined whether the distribution was excessively non-normal, as distributional assumptions are less stringent when using ANOVAs than when using OLS regressions (Schmider et al., 2010; Vercruyssen & Hendrick, 2011). This examination revealed that no variable departed from a normal distribution in such a way as to compromise the conclusions of the F-tests (skewness <2; kurtosis <9).

The comprehension difficulty measures reflected significant differences in the mean ratings for both the *ambiguity* ($M_b = 4.06$, $M_a = 5.12$, $F = 7.32$, $p < 0.01$) and *complexity* subcategories ($M_b = 2.88$, $M_a = 5.65$, $F = 36.32$, $p < 0.001$). Socially desirable items yielded significantly larger mean ratings than the items that were free of social

TABLE 7 Evidence of the validity of the cognitive evaluation procedure

	Mean ratings		
Measure	**Baseline item**	**Altered item**	***F*-test**
Comprehension Difficulty (CD)			
(*Ambiguity*) The statement has several possible meanings.	4.06	5.12	$F(1,33) = 7.32$ **
(*Complexity*) The statement is worded in a complicated manner.	2.88	5.65	$F(1,33) = 36.32$ ***
Social Desirability (SD)			
(*Moralistic*) The statement encourages the respondents to present themselves as "role models" for others.	4.15	5.82	$F(1,33) = 14.11$ ***
(*Egoistic*) The statement encourages the respondents to assert their superiority over others.	4.09	5.59	$F(1,33) = 12.23$ ***
Response Inducement (RI)			
(*Suggested*) The statement contains hints that betray the personal preference of its author.	3.74	5.59	$F(1,33) = 21.61$ ***
(*Implied*) The statement is worded in a way that implies a specific response.	3.77	6.06	$F(1,33) = 39.62$ ***
Knowledge Assumption (KA)			
(*Practical*) The statement assumes that all of the respondents have extensive hands-on experience using the IT.	4.29	5.09	$F(1,33) = 5.38$ **
(*Theoretical*) The statement assumes that all of the respondents are very knowledgeable about IT-related issues.	3.65	5.29	$F(1,33) = 15.39$ ***

Note: $+p < 0.10$; $*p < 0.05$; $**p < 0.01$; $***p < 0.001$.
Abbreviations: CD, comprehension difficulty; KA, knowledge assumption; RI, response inducement; SD, social desirability.

desirability for both the *Moralistic* ($M_b = 4.15$, $M_a = 5.82$, $F = 14.11$, $p < 0.001$) and *Egoistic* ($M_b = 4.09$, $M_a = 5.59$, $F = 12.23$, $p < 0.001$) subcategories. Our manipulation of Response Inducement resulted in significantly larger mean ratings for the *Suggested* ($M_b = 3.74$, $M_a = 5.59$, $F = 21.61$, $p < 0.001$) and *Implied* ($M_b = 3.77$, $M_a = 6.06$, $F = 39.62$, $p < 0.001$) subcategories. Finally, the items that were worded to reflect high degrees of knowledge assumption yielded significantly greater mean ratings when knowledge was of the *practical* ($M_b = 4.29$, $M_a = 5.09$, $F = 5.38$, $p < 0.01$) and *theoretical* ($M_b = 3.65$, $M_a = 5.29$, $F = 15.39$, $p < 0.001$) subcategories.

We find that the cognitive evaluation procedure can discriminate between the baseline and the altered item for each of the eight PIC measures. Although differences in mean ratings vary across the PIC measures, we suspect that these variations are caused by the nature of the manipulations (salient alterations resulting in greater differences in mean ratings) rather than the quality of the measures. We acknowledge that some of the manipulations could produce similar effects on multiple PIC measures. For example, manipulations meant to increase the ambiguity of items could also increase their complexity, which might be reflected in changes in levels for both the Ambiguity and Complexity measures. However, we did not test for this aspect because the impact of these spillover effects is constrained to within one of the four categories. In addition, we argue that this phenomenon is a feature of the nature of the manipulations (and the difficulty of isolating those characteristics at the item wording level) rather than an indication of the deficiency of the measures. Arguably, the measures have strong face validity because each is a direct translation of the definition of the PIC to which it refers.

3.3 | Step 3-evidence of the utility of the procedure (study 2)

Our tests have thus far focused on producing theoretical and statistical evidence for the validity of the method, leaving aside more pragmatic concerns over its usefulness. Step 3 is meant to confront the method with real applications to examine its potential to influence measure development practices in IS research.

3.3.1 | Research design

The research protocol of Study 2 was inspired by research designs that allow us to infer the causality of the method on item design practices. Specifically, Study 2's research design is inspired by a "quasi-experiment" research design (Podsakoff & Podsakoff, 2019). Although our primary intent is to evaluate the influence of the method of item revision practices, we also wanted to verify that people with varying degrees of experience would be able to easily appropriate the method. Accordingly, our research design comprises one within-group factor (treatment: item mean ratings from the method) and one between-group factor (experience: low vs. high). This design is visually depicted in Figure 1.

3.3.2 | Data collection

The implementation of Study 2 required selecting panel members, identifying the scale items to revise, and producing the study material. First, we recruited four panels, with two panels having low levels of experience (Panels 1 and 2) and two panels having high levels of experience (Panels 3 and 4). Accordingly, Panels 1 and 2 comprise three to five Ph.D. candidates in the early stage of their candidacy with some limited exposure to measurement issues but sufficient knowledge to understand the task. Panels 3 and 4 comprise three- to five-scale development experts from various management disciplines (i.e., IS, marketing, and organisational behaviour). All of the experts were trained in psychometric measurement and were familiar conducting field surveys.

		Without Mean Ratings (*control condition*)	With Mean Ratings (*treatment condition*)
PhD Students *(low experience)*	Panel 1	Scale A	Scale B
	Panel 2	Scale B	Scale A
Faculty Members (*high experience*)	Panel 3	Scale A	Scale B
	Panel 4	Scale B	Scale A

FIGURE 1 Quasi-experiment design (Study 2)

Second, we selected two scales published in *MIS Quarterly*—the Perceived Privacy Risk (Wunderlich et al., 2019) and the Exploratory Task Adaptation (Schmitz et al., 2016) scales. We selected these scales because they have high levels of content validity and offer some variance in item wording. We subjected the two scales to the new method and extracted the mean ratings using the cognitive validation procedure validated in Step 2. This input was integrated into the experimental material, and two authors independently carried out the data collection on site. Coordination sessions were set up to train the authors and ensure that the data collection protocol was consistently implemented across the four panels. The support material for Study 2 is provided in Appendix C

The data collection protocol unfolded as follows: 1. definition of the panel objectives; 2. first round of item revision; 3. second round of item revision (treatment condition: with mean ratings); 4. collective panel debriefing; and 5. individual follow-ups. The researchers started by reading the instructions aloud to clarify the purpose of the panel and to outline that the items would be used in subsequent studies. Then, the panel participants were provided the first scale to be revised, which corresponded to the "control condition" (see Tables C1 and C2). When this revision round was over, the second round started by using the second scale. For the second revision round (the "treatment condition"), the participants were given the item ratings obtained using the method in addition to the construct definitions and corresponding scale items (see Tables C3 and C4). When these two activities were completed, the participants were informed about the true purpose of the panel, namely, evaluating the utility of the method, and were given the opportunity to comment on the usefulness of the information (i.e., item mean ratings) that they received during the second round of revision.

3.3.3 | Results and discussion

Study 2 yielded 7 h of video recordings, 9 pages of written testimonies from the post-panel debriefing sessions with the panel participants, and field notes from observations. The data were analysed using a general analytical template that focused on (1) the sequence of activities involved in the item revision process, (2) areas in which the item means ratings contributed to the item review process, and (3) areas of improvement for the method.

We first searched for patterns that indicated that the degree of participants' experience would influence how they appropriate the method but could not find meaningful differences between the panels that comprised Ph.D. students (low experience panels) and the panels that comprised faculty members (high experience panels). We interpreted this finding as an indication that experience was not a barrier to the appropriation of the method. We then contrasted the control and treatment conditions in search of differences that could illuminate the influence of the method of the item design process and its outcome. We initially focused on directly observable metrics (e.g., panel duration, number of modifications, extent of revisions) but could not find significant patterns in the outcome of the item revision activity. Indeed, our observations revealed that extraneous factors such as group dynamics could influence those metrics.

Drawing on these observations, we decided to approach our dataset differently by examining the influence of the method on the item revision process at a more abstract level using a bottom-up analytical approach. This analysis indicated that the method could benefit item review panels in four primary ways: (1) expand the breadth of the

examination, (2) promote a more effective prioritisation of effort, (3) provide faster convergence on item revision decisions, and (4) allow for a stronger focus on problem resolution. The preliminary results of this analysis were discussed within the research team and then formally shared with all of the panel participants, who were invited to comment. The comments were integrated and yielded four categories of benefits, as summarised in Table 8.

3.3.4 | Breadth of examination

One of the most apparent benefits was an increase in the aspects that the panel participants examined when reviewing the items. In the control condition, revisions were primarily focused on improving item readability (i.e., complexity, ambiguity, use of jargon terms). In the treatment condition, the assessment framework served as a guide that invited consideration of a more comprehensive set of issues that could arise at the retrieval, judgement, and response selections stages. Thus, we conclude that using the assessment framework to guide expert reviews tends to improve the range of issues considered during expert panels without compromising the spontaneity and collegial nature of the process.

3.3.5 | Prioritisation of effort

Collectively deciding the items on which the revision effort should concentrate appeared to be a cumbersome activity for the panel members in the control condition. We noticed that the participants engaged with the items in the order that they were presented without prioritising items requiring more attention over the others. When the participants were provided with the item ratings, we observed that they prioritised the items with the highest ratings that significantly deviated from the ratings of the other items (these items were highlighted in bold in the instructions because this is a core feature of the method). The debriefing sessions confirmed that the mean ratings indicated whether a specific problem existed for an item and the magnitude of this problem, which helped prioritise the revision effort. In two of the four panels, item statements that were not highlighted by the new method as potentially problematic were left unchanged.

3.3.6 | Convergent thinking

The assessment of the specific words or grammatical constructions that were problematic often led to intense debate in the control condition. When the item ratings were provided, we observed that the participants converged more quickly on their interpretations of the items that were deemed subject to scrutiny. Differences in the interpretation of the root cause of the problem persisted, but these discussions seemed more targeted and less driven by personal opinions and feelings. According to the panel participants, the provision of item ratings facilitated the revision process and made the activity more time efficient and less mentally taxing.[5]

3.3.7 | Focus on problem resolution

A careful examination of the sequence of activities performed as part of the item revision process revealed that the process could be divided into the following sequence: (1) detection (i.e., identification of problematic items); (2) diagnostic (i.e., analysis of the source of the problem within the item); and (3) resolution (i.e., attenuation or elimination

[5]Note that these conclusions were based on the subjective impressions reported by some of participants and that we could not formally measure them.

TABLE 8 Preliminary evidence of benefits

Benefit category	Observed item revision practices		Exemplary quotes
	Without mean ratings	With mean ratings	
Breadth of examination	Item revisions focus primarily on readability issues	Revision efforts embrace all of the sources of cognitive biases	"In the first session, our feedback was truly based on intuition. Trying to make the item as clear as possible without truly paying attention to all the other potential biases. At the opposite, we were fully driven for the second session since we already knew what the main issues with the items were" (Panel 3, Participant 1)
Prioritisation of effort	Items are handled in a linear way based on the order in which they are presented to the panel	Mean ratings help prioritise revision efforts	"The guidance enables users to focus on precise elements of the items (like social desirability, complexity, induced response, etc.) that require greater attention than others, thus enabling a more precise improvement in the scale items than was possible before" (Panel 2, Participant 1)
Convergent thinking	Participants have different assessments of an item's potential problem	Participants quickly converge on the aspects of an item that may be problematic	"The coordination time among the peer/expert group is longer if you do not have more specific instructions (first session vs. more precise second session) and the team is thinking in many more directions in the first case" (Panel 3, Participant 2)
Focus on problem resolution	Participants expend effort detecting problem items and diagnosing problems with item content	Participants focus effort on problem resolution	Participant 2: "So, the second one has an issue with comprehension difficulty"? Participant 3: "Yes because it is too long, I think, and it is weighted with a 'comma' followed by 'because'" Participant 1: "I was thinking of '[tentative revision]'. It removes the 'because' and the 'comma'. And it is still long enough. Also, I think that it conveys the same meaning as before" (Panel 2, Video Recoding)

of the PICs). In the control condition, we observed that the participants spent a considerable amount of time and effort in detecting problematic items and establishing a common reason for why there was a problem (i.e., convince their peers). This left very little time to discuss possible remedies. Introducing the item ratings facilitated the detection and analysis stages, allowing the participants to focus their effort on discussing possible remedies.

3.4 | Step 4-norms for evaluating cognitive validity in IS (study 3)

Step 3 revealed preliminary evidence of the utility of the method in scale adaptation activities and shed light on areas in which the method could be improved. One such area of improvement was the development of a dedicated index that would help researchers benchmark the performance of their scale against other scales. This need prompted the development of Step 4, whose purpose is to clarify the criteria for interpreting the numerical scores produced by the method.[6]

[6]We also thank the review panel for suggesting expanding our work in this direction.

3.4.1 | Research design

To facilitate the benchmarking of scales against the cognitive validity criteria across studies, the impact of the number of response options used in the data collection procedure must be neutralised. Thus, we introduce the *pic* index (for *problematic item characteristics*), which divides the average ratings obtained across a scale's item by the number of response anchors used in the data collection procedure. The *pic* index is a new statistic for evaluating the performance of measurement items from a cognitive standpoint (see Appendix D for more details about the *pic* index).

Defining the evaluation criteria for interpreting the *pic* index can be done in two primary ways: theoretically, based on the hypothetical distribution of the variable, or empirically, by observing how the variable is distributed to determine the cutoff criteria (Lance et al., 2006). We opted for the second option because our knowledge of the hypothetical distribution of the *pic* indices was limited. This research design is similar to that employed by Colquitt et al. (2019) to determine evaluation criteria for interpreting content validity statistics.

3.4.2 | Data collection

We employed a two-step approach in which we first identified a set of 10 venerable scales that have often been used in IS research and then subjected these scales to the new method using a fresh sample of non-expert judges ($N = 419$). To select the scales, we started with the list of the 10 most common IS constructs following the taxonomy provided by Larsen & Bong (2016) and then retained a highly-cited scale for each of the 10 constructs. Details about the selection process are provided in Appendix D, and the list of the retained scales is provided in Table 9.

The 10 scales were then subjected to the method following the implementation steps outlined in Appendix D. The judges were recruited via the online crowdsourcing platform Prolific. Each judge evaluated the items on one of the eight PIC categories that comprised our framework using the previously validated PIC measures with 7 response anchors, from *1-Strongly disagree* to *7-Strongly agree*. We specified our sample characteristics in such a way that participants were between 22 and 60 years old, reported English as their first language, completed at least an undergraduate degree (BA/BSc/other), and had an approval rate of 90/100 or higher. The items were randomly presented to prevent order effects, and two attention checks were included as per usual when using crowdsourced samples (Oppenheimer et al., 2009). The participants who failed at least one attention check were automatically excluded from the final sample. Our resulting sample ($N = 479$) includes 59% women, and most judges (71%) reported having accumulated more than 5 years of work experience during their lives.

3.4.3 | Results and discussion

To summarise our statistics for the 10 venerable IS scales in our review, we calculated the average *pic* indices on all eight cognitive validity aspects. Thus, our use of *pic* indices reflects averages across a scale's items. Table D1 of Appendix D provides the mean ratings of the 10 scales before they were formatted to the same unit and averaged. Table 10 summarises the distributional properties of the eight indices at the scale level, and Appendix D, Table D2 provides a visualisation of these distributions. A score of 0 would indicate that the scale is free of problematic items, while a score of 1 indicates that the scale contains several problematic items.

These statistics are presented for comparison purposes. Indeed, researchers may be willing to benchmark the cognitive validity of their scales against that of venerable IS scales. If a scale falls within the range allotted to a given *pic* index, the scale demonstrates evidence of cognitive validity. If a scale's *pic* index is below that range, then the scale outperforms venerable IS scales on this index. Scale scoring higher than the upper bound of the range should be subject to scrutiny because it means that the PIC in question may be excessively present in the scale. Generally speaking, we recommend that any scale that produces a score of 0.70 or higher on any *pic* index should be carefully

TABLE 9 List of venerable IS scales retained for the threshold definition

Construct cluster	Corresponding scale (source)	Nb of items
Individual-level usefulness	Perceived usefulness (Davis, 1989)	6
Ease of use	Perceived ease of use (Davis, 1989)	5
Intention to use	Intention to use (Venkatesh et al., 2003)	3
Affect towards technology use	Affect (Compeau et al., 1999)	5
Facilitating conditions	Facilitating conditions (Venkatesh et al., 2012)	4
Individual-level trust: Trustworthiness	Trusting beliefs (McKnight et al., 2002)	11
General concerns about information privacy	Global information privacy concern (Malhotra et al., 2004)	6
Satisfaction with technology	Satisfaction with IT (Wixom & Todd, 2005)	4
Technology self-efficacy	Computer self-efficacy (Compeau & Higgins, 1995)	10
Organisation-level usefulness	Organisational Impact (Gable et al., 2008)	8

Note: construct clusters are based on Larsen and Bong (2016)'s taxonomy.

examined and the appropriate remedies be considered. When *CD – Complexity* is concerned, scores of 0.50 or above would demand careful examination of the readability of the item.

4 | DISCUSSION

The discussion section identifies five implementation steps for executing the method. The advantages and limitations of the method over the other existing ones are then weighted to help scholars decide its relevance with regards to their needs. After discussing the theoretical implications of our work, we delineate several directions for future research.

4.1 | Implementation steps and guidance

To facilitate the appropriation of the method by the IS community, we recommend the following steps: (1) define the research objectives, (2) identify the items to assess, (3) implement the rating task, (4) diagnose potential problems, and (5) implement the appropriate remedies. A summary is offered in Appendix E, Table E1, and relevant illustrations are provided in Tables E2–E5.

4.1.1 | Step 1—Defining the research objectives

Step 1 consists in delineating the objectives of the cognitive validation procedure. Three related considerations should be examined at this step: the purpose of the assessment, its relationship with other assessments, and the origin of the items.

We may consider two main scenarios: in the first scenario, the authors plan to employ an existing scale, but they realise when pretesting the scale with propective respondents that several items have to be reworded to fit the situational context of the study. In this scenario, the cognitive validation technique introduced in this paper is a valuable addition to implement after having adapted the items, and prior to undertaking a more extensive data collection. Indeed, calculating *pic* indices would help flag items whose adaptation constitutes a potential risk.

In the second scenario, the authors have decided to adapt an existing scale before even considering whether the items are worded in a way that is suited to the research setting. In this case, alterations of item wording are primarily

TABLE 10 Descriptives for the *pic* indices

Pic indices	Range	Mean	Median	SD
Comprehension Difficulty (CD)				
Ambiguity	0.55–0.62	0.58	0.57	0.02
Complexity	0.32–0.54	0.38	0.37	0.07
Social Desirability (SD)				
Moralistic	0.60–−0.76	0.67	0.67	0.05
Egoistic	0.53–0.81	0.64	0.64	0.08
Response Inducement (RI)				
Suggested	0.56–0.72	0.64	0.65	0.05
Implied	0.59–0.73	0.66	0.68	0.05
Knowledge Assumption (KA)				
Practical	0.49–0.77	0.65	0.66	0.07
Theoretical	0.56–0.75	0.67	0.68	0.05

Note: Descriptives for each of the eight pic indices are based on a sample of 10 venerable IS scales. The pic indices can take any value between 0 and 1.

driven by theoretical considerations, such as a slight change to the definition of the construct, or a specific instantiation of that construct (see for example the various instances of the computer self-efficacy construct by Marakas et al., 2007). In such situation too, the proposed method can be helpful to determine whether the adapted scale meets cognitive validity requirements.

When considering whether to use this method, researchers should be mindful of its strengths and weaknesses with regard to existing cognitive validation methods. If the intent is to flag items that depart from the rest of the scale due to their linguistic properties, and/or to detect scale-level biases that demand implementing specific procedural or statistical remedies, the method introduced in this paper is particularly adequate. However, if the objective is to gain an understanding of all the potential problems with a research instrument (namely, not just wording issues), then expert reviews, field reviews, or interviews are probably the way forward. Similarly, if the aim is to gain an in-depth understanding of how respondents process items, "think aloud" protocols are more appropriate (Meitinger & Behr, 2016; Miller et al., 2014; Willis, 2004).

4.1.2 | Step 2—Identifying the items to assess

Step 2 considers the items to include in the assessment. We recommend including at least five items in the test so that meaningful conclusions can be drawn. Indeed, the statistical analysis that will be performed once the ratings are collected compares how well a given item performs compared to the other items. In addition, researchers may consider including items from conceptually related constructs, as this allows performing additional comparisons in Step 4. Items of both a formative and reflective nature can be considered for inclusion because the method makes no assumptions regarding whether items from the same scale should correlate with one another or not (Jarvis et al., 2003; Petter et al., 2007).

4.1.3 | Step 3—Implementing the rating procedure

Step 3 is concerned with the practical implementation of the data collection protocol. We consider which PICs shall be assessed, how to design the rating task for maximum validity, and who should perform the assessment.

Although we recommend performing a comprehensive assessment of the cognitive properties of a scale by using each of the eight validated PIC measures, the rating procedure allows to focus on a narrower set of issues. For example, some PICs may not be relevant for some research instruments depending on the construct being measured (e.g., knowledge assumption when items make no presuppositions regarding respondents' knowledge). Besides, authors may prefer to focus on a single cognitive validity issue rather than all of them because they expect it to be particularly salient in their research setting (e.g., social desirability in sensitive contexts).

Researchers should design the rating task by following conventional within-subject experimental design principles in which each individual item represents a distinct treatment. Considerations include preventing rater fatigue (Bradburn, 1978), randomising the appearance of the items (Krosnick & Alwin, 1987), and including instructional or comprehension checks (Oppenheimer et al., 2009). We recommend using fully labelled response options in a Likert-type format with five or seven anchors (Weijters et al., 2010). Tables E2–E4 provide instructions and implementation templates that researchers can directly draw from.

The number of raters should be within the range of 30–50 to satisfy the central limit theorem requirements. Naïve student or adult samples with no particular expertise in questionnaire design should yield highly consistent results given the linguistic nature of the task (Yao et al., 2008). Relying on crowdsourced platform for data collection can significantly lower the turnover time and financial cost of the procedure. Resorting to crowdsourced samples is justified because (a) the inferences that are drawn from the sample are not influenced by the characteristics of the general population, and (b) the task is of linguistic rather than conceptual or theoretical nature (Steelman et al., 2014). Notably, crowdsourced samples have been successfully leveraged at a large scale in similar item assessment tasks. For example, Colquitt et al. (2019) have recruited 6240 MTurkers who they paid $2 to perform a similar item assessment task. Multiple samples may be used to distribute the cognitive effort deployed by raters to perform the assessment, especially if the test involves 20 items or more.

4.1.4 | Step 4—Diagnosing potential problems with a scale

The proposed method allows to flag possible cognitive validity problems at two different levels with any given measure: item and scale level. Therefore, cognitive validity is concerned with two related judgements:

1. "Are PICs that could undermine item performance likely to be present in specific items?"
2. "Are PICs embedded in multiple items in such a way that they may bias the measurement process?"

The test strategy consists in first identifying the items whose mean ratings are greater than the mean ratings of the other items and then testing if these differences are significant at $p < 0.05$ using repeated one-way ANOVAs. When the mean ratings are greater than the Grand Mean (GM), a deviation contrast that compares the mean rating of each level to the mean rating of all the remaining levels is performed. We conclude that the item is weaker than other scale items if the *F*-statistic is statistically significant at $p < 0.05$. A repeated measure one-way ANOVA should be performed so that adjustments are made to the error term (Winer, 1962). This analysis allows researchers to identify items that deviate from the rest of the scale for linguistic reasons.

Once the item-level analysis is performed, we recommend turning to the analysis of the whole scale. Based on the evaluation of 10 venerable IS scales (see Study 3), results within the 0.70–1.00 range would signal a potential problem with the scale for any of the indices. However, if a scale is nearing the 0.70 cut-off point on the *Complexity* dimension, there is ample room for simplifying the wording of its items. Based on our empirical observations, aiming for a score of 0.50 or lower on the *Complexity* dimension would be a more appropriate target. Table E5 introduces a possible way of reporting the results of these tests so that readers and reviewers can appreciate the cognitive properties of a scale.

4.1.5 | Step 5—Implementing the appropriate remedies

In Step 5, items should be revised regarding the problems that have been diagnosed in the previous step. Although the information extracted in Step 4 can help diagnose a problem and specify its nature, it does not provide information about the specific source of the problem or about the remedies that could be implemented to attenuate its effects. Thus, the researcher's judgement, possibly complemented by insights from peers and experts, is fundamental for this step to be carried out successfully. For scale-wide issues, researchers should consider whether the problem can be attenuated by redesigning the scale's items, or if specific procedural and/or statistical remedies should be implemented to mitigate inevitable method biases caused by item wording (Podsakoff et al., 2012).

4.2 | Relative advantages and limitations of the method

Limiting the influence of weakly grounded intuitions and personal experience on the measure development process is perhaps the main contribution of the new method. First, we attenuate the risk that excessive subjectivity contaminates the data generation phase by using large samples of participants. This facilitates the emergence of a consensus that is reflected in the normal distribution of ratings and culminates around the mean score. Second, we minimise the risk that inappropriate decisions are made at the data analysis stage, based on external criteria, for critical item retention decisions (i.e., ANOVAs and F-tests). To enhance the objectivity of the decision-making process, empirically derived thresholds (see Study 3) are also provided. Although the method does not completely do away with subjectivity, it significantly minimises the influence of extraneous factors that typically undermine the reliability of existing cognitive validation techniques (DeMaio & Landreth, 2004; Forsyth et al., 2004; Rothgeb et al., 2007).

The transparency and traceability of research practices is another area that can benefit from the proposed method. Transparency is a critical problem pertaining to management research that can be framed as a "research performance problem" (Aguinis et al., 2018). Improving the performance of theory-testing practices through enhanced transparency is of tremendous importance for IS research. In their review of content validation in IS journals published between 2018 and 2019, Schmitz and Storey (2020) called for a "transition to objectively reproducible methods using quantitative metrics" (p. 43). To enhance methodological transparency, researchers can report figures in the form of ratings along with the results of their statistical tests. It is through enhanced transparency that more appropriate scale adaptation standards will be enacted and espoused by the IS community. This, along with other initiatives, is expected to result in more transparent and consistent measure development practices (Burton-Jones & Lee, 2017; Gregor & Klein, 2014; Schmitz & Storey, 2020).

Despite these advantages, it is of importance to highlight the main tradeoffs that the new method implies relatively to existing cognitive validation methods. On the onset, it should be made clear that the use of ratings inevitably involves significant information loss relative to approaches that elicit people's opinions. Therefore, researchers opting for our method must forego some of the richness of the viewpoints that are typically generated at the item revision stage. For example, cognitive interviews allow for an in-depth analysis of problems through the use of probes in response to something respondents say (Meitinger & Behr, 2016; Miller et al., 2014; Willis, 2004). These probes can reveal buried problems that could be overlooked with a rating approach that looks only at surface item characteristics. Similarly, as indicated in the introduction, experts may disagree in their assessment of potential survey item problems; while confusing at first, these diverging opinions can prove beneficial if properly managed. By favouring the emergence of a homogeneous set of judgements that culminate around the mean, the proposed approach trades richness and diversity for simplicity and consensus.

Another relative shortcoming of the method lies in its relative rigidity in comparison with more loosely defined cognitive validation techniques. For example, no explicit standard exists to determine how many individuals should participate in a field pretest, expert review, or cognitive interview panel. Similarly, there is some ambiguity attached to the term "expert" in review panels: should method or substantive experts be involved? Or else, experts be trained academics or

subject matter experts from practice? (Diamantopoulos et al., 1994). The absence of clearly articulated rules can be beneficial for projects that are more exploratory in nature because researchers can craft protocols that best suit their needs. In contrast, a highly focused approach such as the one introduced in this paper offers limited opportunity for exploration.

4.3 | Theoretical implications

Along with its methodological contributions, this paper makes a theoretical contribution to the measure development literature. In the review section, we have pointed at a possible deficiency in the IS literature by observing that cognitive validity was generally amalgamated with readability issues. We argue that this narrow conception has prevented the IS field from dealing with a wider range of cognitive validity issues. In response to this, we propose a parsimonious conceptual framework that integrates four facets of cognitive validity, namely *comprehension difficulty*, *social desirability*, *response inducement*, and *knowledge assumption*. We also introduce important nuances for each of these facets, by distinguishing "ambiguous" and "complex" wording, "moralistic" and "egoistic" types of item desirability, "suggested" and "implied" wording, or "practical" and "theoretical" types of knowledge assumptions. In Table 4, we make explicit the mechanisms through which insufficient cognitive validity would undermine our ability to effectively measure what we seek to measure (content validity), would be detrimental to psychometric performance (construct validity or reliability), and would lead to inflated substantive relationships (common method effects). By clarifying the cognitive validity concept and its consequences, we contribute to clarifying a key issue pertaining to measures and measurement (Burton-Jones & Lee, 2017).

Further advancing our understanding of the cognitive validity concept requires distinguishing it from the related concept of content validity (Suddaby, 2010). On the one hand, content validity is concerned with the relationship between a construct and its operationalizations in such a way that the two must be congruent for content validity to be satisfied (Schmitz & Storey, 2020). Whether a scale satisfies content validity requirements thus depends on the extent to which items and the construct definition are semantically congruent. This property is typically gauged using sorting or rating procedures where items are matched with a list of constructs (Anderson & Gerbing, 1991; Hinkin & Tracey, 1999; Moore & Benbasat, 1991).

On the other hand, cognitive validity is concerned with the extent to which specific linguistic item characteristics undermine the integrity of the survey response process (Podsakoff et al., 2003; Schwarz, 1999; Sudman et al., 2010; Tourangeau et al., 2000). With cognitive validity, the focus is exclusively on the relationship between the operationalization and the respondent and, more specifically, on the undesirable effects that result from respondents processing the item content. Consequently, while content validity seeks to *maximise the shared meaning* between a construct and its operationalization, cognitive validity aims to *minimise the alteration* resulting from PICs on the survey response process (Figure 2).

This conceptual distinction implies that measures that satisfy cognitive validity requirements are not necessarily valid from a content validity perspective. Indeed, a scale may be perfectly valid from a cognitive standpoint, but it could still capture aspects that are not part of the content domain of the construct (lack of content validity). Conversely, changes in wording driven by content validity considerations could lead to new phrasings that may bias how respondents process item content at the data collection stage (lack of cognitive validity).

This conceptual distinction has two important bearings on how researchers engage with the measure development process. First, given that each concept is concerned with different aspects of measurement, they should be tested using different techniques. Our review of 322 IS studies suggests that cognitive validity and content validity are generally attended simultaneously, which is arguably a suboptimal situation. Moving forward, we recommend uncoupling the cognitive validity and content validity concepts and their assessment. A second implication is that content validity and cognitive validity should be attended iteratively, by considering the recursive effect that one has on the other one. Indeed, modifications of item wording driven by content validity considerations might have adverse consequences on the cognitive validity of the scale, and vice versa.

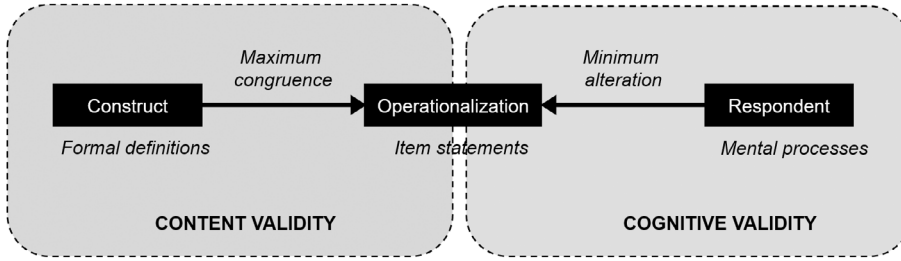


FIGURE 2 Contrasting content validity and cognitive validity

4.4 | Future research directions

Research on the nature and impact of scale adaptation practices is still in its infancy (Compeau et al., 2022; Heggestad et al., 2019) and more work is needed to evaluate the nature of scale adaptations and their impact on measurement in the IS field. In their systematic review, Heggestad et al. (2019) found that authors frequently — up to almost 50% of the time — altered the wording of scale items, and that these changes are usually fairly substantive. Replicating such analysis in the context of IS would refine our understanding of the frequency of scale adaptations as well as the nature of the changes that are routinely made to survey items. Although the review of 54 scales that was conducted in this paper is a step in this direction, it should be extended to advance our understanding of measure development practices. For instance, Heggestad et al. (2019) argue that the progressive shortening of psychometric scales is an important threat to content validity. Indeed, items representing more peripheral aspects of the content domain are at risk of being excluded from successive adaptations despite their importance in representing relevant facets of a construct (Stanton et al., 2002). Therefore, although our analysis suggests that item wording is a critical issue in empirical IS research, other concerns may surface from further investigations of scale adaptation practices.

Although it is widely recognised that method effects are detrimental to IS research (Burton-Jones, 2009; Sharma et al., 2009), we are still unsure about the extent to which specific effects produce by item characteristics represent a threat to the validity of empirical IS research. In fact, recent developments in the measurement literature suggest that method effects should not be approached from a broad perspective of the method but rather from the microscopic vantage point of the individual measure (Spector et al., 2019). Using this approach, Kwak et al. (2019) were able to quantify the extent to which SD bias can impact IS research, and to demonstrate that this type of response effects is heightened by negative and sensitive contexts. Their approach to item characteristic effects could be extended to the other sources outlined in this paper, namely, CD, RI, and KA. To complement these insights, IS researchers could investigate the specific mechanisms that lead to these effects by purposively manipulating item characteristics to isolate the influence of these characteristics on response patterns (Kam & Fan, 2020; Kuncel & Tellegen, 2009; Spector et al., 1997). The procedure introduced in Table 6 can serve this purpose.

Finally, recent advances in linguistic computing have resulted in the development of language-based algorithms that can directly contribute to improving measurement practices in management research (Gefen & Larsen, 2017; Larsen & Bong, 2016; Nimon et al., 2016; Rosenbusch et al., 2020). We view these developments, supported by NLP and machine learning (ML) techniques, as an opportunity to extend the groundwork laid by this paper. Researchers could leverage the evaluation framework introduced in this paper to inform the development of a more detailed ontology geared towards improving cognitive validity (see Lukyanenko et al., 2019, for a general ontology for research validity). In addition, our method can generate data that could then be used to train supervised ML models to detect problematic items (McDaniel & Storey, 2019). Indeed, algorithms that rely on ML often resort to crowdsourced samples to generate training data that are then used for algorithm calibration (Mohammad, 2016). Automating the cognitive validity assessment process would lower the costs associated with this step because this

would eschew the need to resort to human judges. This would also contribute to refining the norms for gauging cognitive validity that this paper has introduced in Study 3.

5 | CONCLUSION

We started our work by problematizing some of the taken-for-granted assumptions pertaining to scale adaptation practices. Evidence suggested that considerations pertaining to the cognitive validity of measures is partially attended to at the measure development stage. We argue that insufficient attention to the cognitive validity of adapted IS scales can undermine the ability of these scales to measure the constructs they purport to measure. This can also have downstream effects on the psychometric performance of individual items at the data collection stage or produce artifactual covariances that inflate substantive relationships. Our response to this problematic situation is two-fold: we first define the cognitive validity concept and provide an organising framework. We then introduce a method to gauge the extent to which PICs manifest in measurement scales.

It is our hope that this work will encourage positivist, quantitative IS researchers to examine more systematically whether their measures are worded in a way that minimise the salience of PICs. We believe that by making the execution and interpretation of cognitive validity assessments clearer, this work will pave the way to more robust, transparent, and consistent measure development practices. The need to pay extra attention to item wording quality is becoming increasingly critical as management research tends to rely on fewer and fewer items to measure latent constructs (Cortina et al., 2020; Heggestad et al., 2019).

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Jean-Charles Pillet https://orcid.org/0000-0002-7247-2408
Kevin D. Carillo https://orcid.org/0000-0002-9714-1621
Claudio Vitari https://orcid.org/0000-0002-2191-5205
Federico Pigni https://orcid.org/0000-0002-9826-3815

REFERENCES

Agarwal, R., & Prasad, J. (1998). A conceptual and operational definition of personal innovativeness in the domain of information technology. *Information Systems Research*, *9*(2), 204–215. https://doi.org/10.1287/isre.9.2.204

Agresti, A. (2017). *Statistical methods for the social sciences*. Pearson.

Aguinis, H., Ramani, R. S., & Alabduljader, N. (2018). What you see is what you get? Enhancing methodological transparency in management research. *Academy of Management Annals*, *12*(1), 83–110. https://doi.org/10.5465/annals.2016.0011

Al-Natour, S., Cavusoglu, H., Benbasat, I., & Aleem, U. (2020). An empirical investigation of the antecedents and consequences of privacy uncertainty in the context of mobile apps. *Information Systems Research*, *7*(12), 821-861. https://doi.org/10.1287/isre.2020.0931

Alvesson, M., & Sandberg, J. (2011). Generating research questions through problematization. *Academy of Management Review*, *36*(2), 247–271. http://www.jstor.org/stable/41318000

Anderson, J. C., & Gerbing, D. W. (1991). Predicting the performance of measures in a confirmatory factor analysis with a pretest assessment of their substantive validities. *Journal of Applied Psychology*, *76*(5), 732–740. https://doi.org/10.1037/0021-9010.76.5.732

Bagozzi, R. P. (2011). Measurement and meaning in information systems and organizational research: Methodological and philosophical foundations. *MIS Quarterly*, *35*(2), 261–292. https://doi.org/10.2307/23044044

Baskerville, R. L., & Myers, M. D. (2002). Information systems as a reference discipline. *MIS Quarterly*, *26(1)*, 1– 14. https://www.jstor.org/stable/4132338

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. https://doi.org/10.1509/jmkr.38.2.143.18840

Bélanger, F., & Crossler, R. E. (2011). Privacy in the digital age: A review of information privacy research in information systems. *MIS Quarterly*, *35*(4), 1017–1041. https://doi.org/10.2307/41409971

Benlian, A. (2020). A Daily field investigation of technology-driven spillovers from work to home. *MIS Quarterly*, *44*(3), 1259–1300. https://doi.org/10.25300/MISQ/2020/14911

Bhattacherjee, A. (2001). Understanding information systems continuance: An expectation-confirmation model. *MIS Quarterly*, *25*(3), 351–370. https://doi.org/10.2307/3250921

Blair, E., & Burton, S. (1987). Cognitive processes used by survey respondents to answer behavioral frequency questions. *Journal of Consumer Research*, *14*(2), 280–288. https://doi.org/10.1086/209112

Boudreau, M.-C., Gefen, D., & Straub, D. W. (2001). Validation in information systems research: A state-of-the-art assessment. *MIS Quarterly*, *25*(1), 1–16. https://doi.org/10.2307/3250956

Bradburn, N. (1978). Respondent burden. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 35–40. http://www.asasrms.org/Proceedings/y1978f.html

Bradburn, N., Sudman, S., & Wansink, B. (2004). *Asking questions: A practical guide to questionnaire design*. Jossey-Bass.

Burton-Jones, A. (2009). Minimizing method bias through programmatic research. *MIS Quarterly*, *33*(3), 445–471. https://doi.org/10.2307/20650304

Burton-Jones, A., & Lee, A. S. (2017). Thinking about measures and measurement in positivist research: A proposal for refocusing on fundamentals. *Information Systems Research*, *28*(3), 451–467. https://doi.org/10.1287/isre.2017.0704

Burton-Jones, A., & Straub, D. W. (2004). Minimizing method variance in measures of system usage. *Southern Association for Information Systems*, *1979*, 336–342.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Caplan, S. E. (2002). Problematic Internet use and psychosocial well-being: Development of a theory-based cognitive–behavioral measurement instrument. *Computers in Human Behavior*, *18*(5), 553–575. https://doi.org/10.1016/S0747-5632(02)00004-3

Carver, C. S. (1997). You want to measure coping but your protocol' too long: Consider the brief cope. *International Journal of Behavioral Medicine*, *4*(1), 92–100. https://doi.org/10.1207/s15327558ijbm0401_6

Cassiman, B., Colombo, M. G., Garrone, P., & Veugelers, R. (2005). The impact of M&A on the R&D process: An empirical analysis of the role of technological- and market-relatedness. *Research Policy*, *34*(2), 195–220. https://doi.org/10.1016/j.respol.2005.01.002

Chan, H., Wan, L. C., & Sin, L. Y. M. (2009). The contrasting effects of culture on consumer tolerance: Interpersonal face and impersonal fate. *Journal of Consumer Research*, *36*(2), 292–304. https://doi.org/10.1086/597329.

Charlton, J. P. (2002). A factor-analytic investigation of computer 'addiction' and engagement. *British Journal of Psychology*, *93*(3), 329–344. https://doi.org/10.1348/000712602760146242

Chatterjee, S., & Davison, R. M. (2021). The need for compelling problematisation in research: The prevalence of the gap-spotting approach and its limitations. *Information Systems Journal*, *31*(2), 227–230. https://doi.org/10.1111/isj.12316

Chen, A., & Karahanna, E. (2018). Life interrupted: The effects of technology-mediated work interruptions on work and non-work outcomes. *MIS Quarterly: Management Information Systems*, *42*(4), 1023–1042. https://doi.org/10.25300/MISQ/2018/13631

Chen, D. Q., Preston, D. S., & Xia, W. (2010). Antecedents and effects of CIO supply-side and demand-side leadership: A staged maturity model. *Journal of Management Information Systems*, *27*(1), 231–272. https://doi.org/10.2753/MIS0742-1222270110

Chengalur-smith, I. S., Nevo, S., & Demertzoglou, P. (2010). An empirical analysis of the business value of open source infrastructure technologies. *Journal of the Association for Information Systems*, *11*(11), 708–729. https://doi.org/10.17705/1jais.00242

Chin, W. W., Johnson, N., & Schwarz, A. (2008). A fast form approach to measuring technology acceptance and other constructs. *MIS Quarterly*, *32*(4), 687–703. https://doi.org/10.2307/25148867

Clifton, C., & Duffy, S. A. (2001). Sentence and text comprehension: Roles of linguistic structure. *Annual Review of Psychology*, *52*(1), 167–196. https://doi.org/10.1146/annurev.psych.52.1.167

Collopy, F. (1996). Biases in retrospective self-reports of time use: An empirical study of computer users. *Management Science*, *42*(5), 758–767. https://doi.org/10.1287/mnsc.42.5.758

Colquitt, J. A., Sabey, T. B., Rodell, J. B., & Hill, E. T. (2019). Content validation guidelines: Evaluation criteria for definitional correspondence and definitional distinctiveness. *Journal of Applied Psychology*, *104*(10), 1243–1265. https://doi.org/10.1037/apl0000406

Compeau, D. R., Correia, J., & Thatcher, J. B. (2022). When constructs become obsolete: A systematic approach to evaluating and updating constructs for information systems research. *MIS Quarterly*, *46*(2), 679–712. https://doi.org/10.25300/MISQ/2022/15516

Compeau, D. R., & Higgins, C. A. (1995). Computer self-efficacy: Development of a measure and initial test. *MIS Quarterly*, *19*(2), 189–211. https://doi.org/10.2307/249688

Compeau, D. R., Higgins, C. A., & Huff, S. (1999). Social cognitive theory and individual reactions to computing technology: A longitudinal study. *MIS Quarterly*, *23*, 145–158. https://doi.org/10.2307/249749

Cortina, J. M., Sheng, Z., Keener, S. K., Keeler, K. R., Grubb, L. K., Schmitt, N., Tonidandel, S., Summerville, K. M., Heggestad, E. D., & Banks, G. C. (2020). From alpha to omega and beyond! A look at the past, present, and (possible) future of psychometric soundness in the journal of applied psychology. *Journal of Applied Psychology*, *105*(12), 1351. https://doi.org/10.1037/apl0000815

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*(4), 349–354. https://doi.org/10.1037/h0047358

Cui, T., Tong, Y., Teo, H.-H., & Li, J. (2020). Managing knowledge distance: IT-enabled inter-firm knowledge capabilities in collaborative innovation. *Journal of Management Information Systems*, *37*(1), 217–250. https://doi.org/10.1080/07421222.2019.1705504

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319–340. https://doi.org/10.2307/249008

de Reuver, M., & Bouwman, H. (2015). Dealing with self-report bias in mobile internet acceptance and usage studies. *Information & Management*, *52*(3), 287–294. https://doi.org/10.1016/j.im.2014.12.002

DeMaio, T. J., & Landreth, A. (2004). Do different cognitive interview techniques produce different results? In *Methods for testing and evaluating survey questionnaires* (pp. 89–108). Wiley Online Library. https://doi.org/10.1002/0471654728

Diamantopoulos, A., Reynolds, N., & Schlegelmilch, B. (1994). Pretesting in questionnaire design: The impact of respondent characteristics on error detection. *Journal of the Market Research Society*, *36*(4), 1–15. https://doi.org/10.1177/147078539403600402

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: The tailored design method*. John Wiley & Sons, Ltd.

Dimoka, A., Hong, Y., & Pavlou, P. A. (2012). On product uncertainty in online markets: Theory and evidence. *MIS Quarterly*, *36*(2), 395–426. https://doi.org/10.2307/41703461

Doty, D. H., & Glick, W. H. (1998). Common methods bias: Does common methods variance really bias results? *Organizational Research Methods*, *1*(4), 374–406. https://doi.org/10.1177/109442819814002

Forsyth, B., Rothgeb, J. M., & Willis, G. B. (2004). Does pretesting make a difference? An experimental test. In *Methods for testing and evaluating survey questionnaires* (pp. 525–546). Wiley Online Library.

Fowler, F. J. (1992). How unclear terms affect survey data. *Public Opinion Quarterly*, *56*(2), 218–231. https://doi.org/10.1086/269312

Gable, G. G., Sedera, D., & Chan, T. (2008). Re-conceptualizing information system success: The IS-impact measurement model. *Journal of the Association for Information Systems*, *9*(7), 377–408. https://doi.org/10.17705/1jais.00164

Galliers, R. D. (2003). Change as crisis or growth? Toward a trans-disciplinary view of information systems as a field of study: A response to Benbasat and Zmud's call for returning to the IT artifact. *Journal of the Association for Information Systems*, *4*(1), 13–352. https://doi.org/10.17705/1jais.00040

Gefen, D., Endicott, J. E., Fresneda, J. E., Miller, J., & Larsen, K. R. (2017). A guide to text analysis with latent semantic analysis in R with annotated code: Studying online reviews and the stack exchange community. *Communications of the Association for Information Systems*, *41*(1), 21–496. https://doi.org/10.17705/1CAIS.04121

Gefen, D., & Larsen, K. R. (2017). Controlling for lexical closeness in survey research: A demonstration on the technology acceptance model. *Journal of the Association for Information Systems*, *18*(10), 727–757. https://doi.org/10.17705/1jais.00469

Gregor, S., & Klein, G. (2014). Eight obstacles to overcome in the theory testing genre. *Journal of the Association for Information Systems*, *15*(11), 1–14. https://doi.org/10.17705/1jais.00382

Hardy, B., & Ford, L. R. (2014). It's not me, it's you: Miscomprehension in surveys. *Organizational Research Methods*, *17*(2), 138–162. https://doi.org/10.1177/1094428113520185

Heggestad, E. D., Scheaf, D. J., Banks, G. C., Monroe Hausfeld, M., Tonidandel, S., & Williams, E. B. (2019). Scale adaptation in organizational science research: A review and best-practice recommendations. *Journal of Management*, *45*(6), 2596–2627. https://doi.org/10.1177/0149206319850280

Hinkin, T. R., & Tracey, J. B. (1999). An analysis of variance approach to content validation. *Organizational Research Methods*, *2*(2), 175–186. https://doi.org/10.1177/109442819922004

Hoehle, H., & Venkatesh, V. (2015). Mobile application usability: Conceptualization and instrument development. *MIS Quarterly*, *39*(2), 435–472. https://www.jstor.org/stable/26628361

Holt, D. T., Armenakis, A. A., Feild, H. S., & Harris, S. G. (2007). Readiness for organizational change: The systematic development of a scale. *The Journal of Applied Behavioral Science*, *43*(2), 232–241 244-245, 247-249, 251-255. https://doi.org/10.1177/0021886306295295

Hsu, M. H., & Chang, C. M. (2014). Examining interpersonal trust as a facilitator and uncertainty as an inhibitor of intra-organisational knowledge sharing. *Information Systems Journal*, *24*(2), 119–142. https://doi.org/10.1111/isj.12000

Jarvis, C. B., Mackenzie, S. B., & Podsakoff, P. M. (2003). A critical review of construct indicators and measurement model Misspecication in marketing and consumer research. *Journal of Consumer Research*, *30*(2), 199–218. https://doi.org/10.1086/376806

Jen-Hwa Hu, P., Han-Fen, H., & Xiao, F. (2017). Examining the mediating roles of cognitive load and performance outcomes in user satisfaction with a website: A field quasi-experiment. *MIS Quarterly*, *41*(3), 975–988. https://doi.org/10.25300/MISQ/2017/41.3.

Jia, R., & Reich, B. H. (2013). IT service climate, antecedents and IT service quality outcomes: Some initial evidence. *Journal of Strategic Information Systems*, *22*(1), 51–69. https://doi.org/10.1016/j.jsis.2012.10.001

Johnson, B., & Christensen, L. (2008). *Educational research: Quantitative, qualitative, and mixed approaches*. Sage Publications.

Johnson, J. A. (1986). Ambiguity, subtlety, and validity of items in the California Psychological Inventory. In *57th annual meeting of the Eastern Psychological Association* (pp. 1–12). Eastern Psychological Association.

Johnson, J. A. (2004). The impact of item characteristics on item and scale validity. *Multivariate Behavioral Research*, *39*(2), 303–328. https://doi.org/10.1207/s15327906mbr3902_6

Jokić-Begić, N., Mikac, U., Čuržik, D., & Sangster Jokić, C. (2019). The development and validation of the short cyberchondria scale (SCS). *Journal of Psychopathology and Behavioral Assessment*, *41*(4), 662–676. https://doi.org/10.1007/s10862-019-09744-z

Kam, C. C. S., & Fan, X. (2020). Investigating response heterogeneity in the context of positively and negatively worded items by using factor mixture modeling. *Organizational Research Methods*, *23*(2), 322–341. https://doi.org/10.1177/1094428118790371

Keen, P. G. W. (1980). *MIS research: Reference disciplines and a cumulative tradition*. Proceedings of the International Conference of Infromation Systems. Association for Information Systems.

Kirsch, L. J., Sambamurthy, V., Ko, D.-G., & Purvis, R. L. (2002). Controlling information systems development projects: The view from the client. *Management Science*, *48*(4), 484–498. https://doi.org/10.1287/mnsc.48.4.484.204

Knowles, E. S., & Condon, C. A. (1999). Why people say" yes": A dual-process theory of acquiescence. *Journal of Personality and Social Psychology*, *77*(2), 379–386. https://psycnet.apa.org/doi/10.1037/0022-3514.77.2.379

Kossek, E. E., Colquitt, J. A., & Noe, R. A. (2001). Caregiving decisions, well-being, and performance: The effects of place and provider as a function of dependent type and work-family climates. *Academy of Management Journal*, *44*(1), 29–44. https://doi.org/10.5465/3069335

Krosnick, J. A. (2018). Improving question design to maximize reliability and validity. In *The Palgrave handbook of survey research* (pp. 95–101). Springer.

Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly*, *51*(2), 201–219. https://doi.org/10.1086/269029

Kuem, J., Ray, S., Hsu, P.-F., & Khansa, L. (2020). Smartphone addiction and conflict: An incentive-sensitisation perspective of addiction for information systems. *European Journal of Information Systems*, *30*(4), 1–22. https://doi.org/10.1080/0960085X.2020.1803154

Kuncel, N., & Tellegen, A. (2009). A conceptual and empirical reexamination of the measurement of the social desirability of items. *Personnel Psychology*, *62*(2), 201–228. https://doi.org/10.1111/j.1744-6570.2009.01136.x.

Kwak, D. H., Holtkamp, P., & Kim, S. S. (2019). Measuring and controlling social desirability bias: Applications in information systems research. *Journal of the Association for Information Systems*, *20*(4), 317–345. https://doi.org/10.17005/1jais.00537

Laato, S., Islam, A. K. M. N., Islam, M. N., & Whelan, E. (2020). What drives unverified information sharing and cyberchondria during the COVID-19 pandemic? *European Journal of Information Systems*, *29*(3), 288–305. https://doi.org/10.1080/0960085X.2020.1770632

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods*, *9*(2), 202–220. https://doi.org/10.1177/1094428105284919

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (2013). *Handbook of latent semantic analysis*. Psychology Press.

Larsen, K. R., & Bong, C. H. (2016). A tool for addressing construct identity in literature reviews and meta-analyses. *MIS Quarterly*, *40*(3), 529–551. https://www.jstor.org/stable/26629026

Li, Y. (2012). Theories in online information privacy research: A critical review and an integrated framework. *Decision Support Systems*, *54*(1), 471–481. https://doi.org/10.1016/j.dss.2012.06.010

Lowry, P. B., Gaskin, J., Twyman, N., Hammer, B., & Roberts, T. (2012). Taking 'fun and games' seriously: proposing the hedonic-motivation system adoption model (HMSAM) (SSRN Scholarly Paper No. 2177442). https://papers.ssrn.com/abstract=2177442

Lukyanenko, R., Larsen, K. R., Parsons, J., & Gefen, D. (2019). Toward creating a general ontology for research validity. International Conference on Conceptual Modeling.

MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Construct measurement and validation procedures in MIS and behavioral research: Integrating new and existing techniques. *MIS Quarterly*, *35*(2), 293–334. https://doi.org/10.2307/23044045

Malhotra, N. K., Kim, S. S., & Agarwal, J. (2004). Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *Information Systems Research*, *15*(4), 336–355. https://doi.org/10.1287/isre.1040.0032

Mäntymäki, M., Islam, A. K. M. N., & Benbasat, I. (2020). What drives subscribing to premium in freemium services? A consumer value-based view of differences between upgrading to and staying with premium. *Information Systems Journal*, *30*(2), 295–333. https://doi.org/10.1111/isj.12262

Marakas, G. M., Johnson, R. D., & Clay, P. F. (2007). The evolving nature of the computer self-efficacy construct: An empirical investigation of measurement construction, validity, reliability and stability over time. *Journal of the Association for Information Systems*, *8*(1), 16–46. https://doi.org/10.17705/1jais.00112

McDaniel, M., & Storey, V. C. (2019). Evaluating domain ontologies: Clarification, classification, and challenges. *ACM Computing Surveys*, *52*(4), 1–44. https://doi.org/10.1145/3329124

McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research*, *13*(3), 334–359. https://doi.org/10.1287/isre.13.3.334.81

Meitinger, K., & Behr, D. (2016). Comparing cognitive interviewing and online probing: Do they find similar results? *Field Methods*, *28*(4), 363–380. https://doi.org/10.1177/1525822X15625866

Miller, K., Chepp, V., Willson, S., & Padilla, J.-L. L. (2014). *Cognitive interviewing methodology*. John Wiley & Sons.

Milne, S., Orbell, S., & Sheeran, P. (2002). Combining motivational and volitional interventions to promote exercise participation: Protection motivation theory and implementation intentions. *British Journal of Health Psychology*, *7*(2), 163–184. https://doi.org/10.1348/135910702169420

Mohammad, S. M. (2016). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In H. L. Meiselman (Ed.), *Emotion Measurement* (pp. 201–237). Woodhead Publishing. https://doi.org/10.1016/B978-0-08-100508-8.00009-6

Moore, G. C., & Benbasat, I. (1991). Development of an instrument to measure the perceptions of adopting an information technology innovation. *Information Systems Research*, *2*(3), 192–223. https://doi.org/10.1287/isre.2.3.192

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*, 1321–1339. https://doi.org/10.1037/0021-9010.91.6.1321

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology*, *15*(3), 263–280. https://doi.org/10.1002/ejsp.2420150303

Newman, D. A., Harrison, D. A., Carpenter, N. C., & Rariden, S. M. (2016). Construct mixology: Forming new management constructs by combining old ones. *Academy of Management Annals*, *10*(1), 943–995. https://doi.org/10.5465/19416520.2016.1161965

Nimon, K., Shuck, B., & Zigarmi, D. (2016). Construct overlap between employee engagement and job satisfaction: A function of semantic equivalence? *Journal of Happiness Studies*, *17*(3), 1149–1171. https://doi.org/10.1007/s10902-015-9636-6

Nunnally, J. C., & Bernstein, I. (1994). *Psychometric theory*. McGraw-Hill.

Okazaki, S., & Mendez, F. (2013). Exploring convenience in mobile commerce: Moderating effects of gender. *Computers in Human Behavior*, *29*(3), 1234–1242. https://doi.org/10.1016/j.chb.2012.10.019

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867–872. https://doi.org/10.1016/j.jesp.2009.03.009

Paulhus, D. L. (2001). Socially desirable responding: The evolution of a construct. In *The role of constructs in psychological and educational measurement* (pp. 49–69). Taylor & Francis Group.

Pavlou, P. A. (2011). State of the information privacy literature: Where are we now and where should we go? *MIS Quarterly*, *35*(4), 977–988. https://doi.org/10.2307/41409969

Pavlou, P. A., & El Sawy, O. A. (2006). From IT leveraging competence to competitive advantage in turbulent environments: The case of new product development. *Information Systems Research*, *17*(3), 198–227. https://doi.org/10.1287/isre.1060.0094

Payne, S. G. (1951). *The art of asking questions*. Princeton University Press.

Petter, S., Straub, D. W., & Rai, A. (2007). Specifying formative constructs in information systems research. *MIS Quarterly*, *31*(4), 623–656. https://doi.org/10.2307/25148814

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. https://doi.org/10.1037/0021-9010.88.5.879

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology*, *63*(1), 539–569. https://dx.doi.org/10.1146/annurev-psych-120710-100452

Podsakoff, P. M., & Podsakoff, N. P. (2019). Experimental designs in management and leadership research: Strengths, limitations, and recommendations for improving publishability. *The Leadership Quarterly*, *30*(1), 11–33. https://doi.org/10.1016/j.leaqua.2018.11.002

Qahri-Saremi, H., & Turel, O. (2020). Ambivalence and Coping Responses in Post-Adoptive Information Systems Use. *Journal of Management Information Systems*, *37*(3), 820–848. https://doi.org/10.1080/07421222.2020.1790193

Rai, A. (2020). Editor's comments: Proactively attending to uncertainty in IS research, *MIS Quarterly*, *44*(1), 3–7.

Remus, U., Wiener, M., Saunders, C., & Mähring, M. (2020). The impact of control styles and control modes on individual-level outcomes: A first test of the integrated IS project control theory. *European Journal of Information Systems*, *29*(2), 134–152. https://doi.org/10.1080/0960085X.2020.1718008

Rosenbusch, H., Wanders, F., & Pit, I. L. (2020). The semantic scale network: An online tool to detect semantic overlap of psychological scales and prevent scale redundancies. *Psychological Methods*, *25*(3), 380–392. https://psycnet.apa.org/doi/10.1037/met0000244

Rothgeb, J., Willis, G. B., & Forsyth, B. (2007). Questionnaire pretesting methods: Do different techniques and different organizations produce similar results? *Bulletin of Sociological Methodology*, *96*(1), 5–31. https://doi.org/10.1177/075910630709600103

Ryu, C., Kim, Y. J., Chaudhury, A., & Rao, H. R. (2005). Knowledge acquisition via three learning processes in enterprise information portals: Learning-by-investment, learning-by-doing, and learning-from-others. *MIS Quartely*, *29*(2), 245–278. https://doi.org/10.2307/25148679

Schmider, E., Ziegler, M., Danay, E., Beyer, L., & Bühner, M. (2010). Is it really robust? *Methodology*, *6*, 147–151. https://doi.org/10.1027/1614-2241/a000016

Schmitz, K. W., & Storey, V. C. (2020). Empirical test guidelines for content validity: Wash, rinse, and repeat until clean. *Communications of AIS*, *47*, 1–65. https://doi.org/10.17705/1CAIS.04736

Schmitz, K. W., Teng, J. T. C., & Webb, K. J. (2016). Capturing the complexity of malleable IT use: Adaptive structuration theory for individuals availability. *MIS Quarterly*, *40*(3), 663–686. https://www.jstor.org/stable/26629032

Schriesheim, C. A., Powers, K. J., Scandura, T. A., Gardiner, C. C., & Lankau, M. J. (1993). Improving construct measurement in management research: Comments and a quantitative approach for assessing the theoretical content adequacy of paper-and-pencil survey-type instruments. *Journal of Management*, *19*(2), 385–417. https://doi.org/10.1016/0149-2063(93)90058-U

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, *54*(2), 93–105. https://psycnet.apa.org/doi/10.1037/0003-066X.54.2.93

Schwarz, N. (2007). Cognitive aspects of survey methodology. *Applied Cognitive Psychology*, *21*(2), 277–287. https://doi.org/10.1002/acp.1340

Schwarz, N., & Sudman, S. E. (1996). *Answering questions: Methodology for determining cognitive and communicative processes in survey research*. Jossey-Bass.

Segars, A. H., & Grover, V. (1999). Profiles of strategic information systems planning. *Information Systems Research*, *10*(3), 199–232. https://doi.org/10.1287/isre.10.3.199

Seiders, K., Voss, G. B., Godfrey, A. L., & Grewal, D. (2007). SERVCON: Development and validation of a multidimensional service convenience scale. *Journal of the Academy of Marketing Science*, *35*(1), 144–156. https://doi.org/10.1007/s11747-006-0001-5

Sharma, R., Yetton, P., & Crawford, J. (2009). Estimating the effect of common method variance: The method-method pair technique with an illustration from TAM research. *MIS Quarterly*, *33*(3), 473–490. https://doi.org/10.2307/20650305

Silic, M., & Lowry, P. B. (2020). Using design-science based gamification to improve organizational security training and compliance. *Journal of Management Information Systems*, *37*(1), 129–161. https://doi.org/10.1080/07421222.2019.1705512

Snell, S. A. (1992). Control theory in strategic human resource management: The mediating effect of administrative information. *Academy of Management Journal*, *35*(2), 292–327. https://doi.org/10.5465/256375

Spector, P. E., Rosen, C. C., Richardson, H. A., Williams, L. J., & Johnson, R. E. (2019). A new perspective on method variance: A measure-centric approach. *Journal of Management*, *45*(3), 855–880. https://doi.org/10.1177/0149206316687295

Spector, P. E., Van Katwyk, P. T., Brannick, M. T., & Chen, P. Y. (1997). When two factors Don't reflect two constructs: How item characteristics can produce Artifactual factors. *Journal of Management*, *23*(5), 659–677. https://doi.org/10.1016/S0149-2063(97)90020-9

Srivastava, S. C., Chandra, S., & Shirish, A. (2015). Technostress creators and job outcomes: Theorising the moderating influence of personality traits. *Information Systems Journal*, *25*(4), 355–401. https://doi.org/10.1111/isj.12067

Stalnaker, R. (1977). Pragmatic presuppositions. In *Proceedings of the Texas conference on performatives, presuppositions, and Implicatures* (pp. 135–148). Center for Applied Linguistics.

Stanton, J. M., Sinar, E. F., Balzer, W. K., & Smith, P. C. (2002). Issues and strategies for reducing the length of self-report scales. *Personnel Psychology*, *55*(1), 167–194. https://doi.org/10.1111/j.1744-6570.2002.tb00108.x

Steelman, Z. R., Hammer, B. I., & Limayem, M. (2014). Data collection in the digital age: Innovative alternatives to student samples. *MIS Quarterly*, *38*(2), 355–378. https://www.jstor.org/stable/26634930

Steenkamp, J. B. E. M., de Jong, M. G., & Baumgartner, H. (2010). Socially desirable response tendencies in survey research. *Journal of Marketing Research*, *47*(2), 199–214. https://doi.org/10.1509/jmkr.47.2.199

Straub, D. W. (1989). Validating instruments in MIS research. *MIS Quarterly*, *13*(2), 147–169. https://www.jstor.org/stable/248922

Straub, D. W., & Gefen, D. (2004). Validation guidelines for is positivist. *Communications of the Association for Information Systems*, *13*(1), 380–427. https://doi.org/10.17705/1CAIS.01324

Suddaby, R. (2010). Editor's comments: Construct clarity in theories of management and organization. *Academy of Management Review*, *35*(3), 346–357. https://doi.org/10.5465/amr.35.3.zok346

Sudman, S., Bradburn, N. M., & Schwarz, N. (2010). *Thinking about answers: The application of cognitive processes to survey methodology*. Jossey-Bass Publishers, Inc.

Sun, H. (2012). Understanding user revisions when using information system features: Adaptive system use and triggers. *MIS Quarterly*, *36*(2), 453–478. https://www.jstor.org/stable/41703463

Tams, S., Ahuja, M., Thatcher, J., & Grover, V. (2020). Worker stress in the age of mobile technology: The combined effects of perceived interruption overload and worker control. *The Journal of Strategic Information Systems*, *29*(1), 101595. https://doi.org/10.1016/j.jsis.2020.101595

Tarafdar, M., Maier, C., Laumer, S., & Weitzel, T. (2020). Explaining the link between technostress and technology addiction for social networking sites: A study of distraction as a coping behavior. *Information Systems Journal*, *30*(1), 96–124. https://doi.org/10.1111/isj.12253

Tourangeau, R. (2018). The survey response process from a cognitive viewpoint. *Quality Assurance in Education.*, *26*, 169–181.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press.

Trenz, M., Veit, D. J., & Tan, C.-W. (2020). Disentangling the impact of Omni Channel integration on consumer behavior in integrated sales channels. *MIS Quarterly*, *44*(3), 1207–1258. https://doi.org/10.25300/MISQ/2020/14121

Turel, O., & Serenko, A. (2012). The benefits and dangers of enjoyment with social networking websites. *European Journal of Information Systems*, *21*(5), 512–528. https://doi.org/10.1057/ejis.2012.1

Vannette, D. L., & Krosnick, J. A. (2018). *The Palgrave handbook of survey research*. Springer.

Venkatesh, V., Davis, F., Cheung, C., & Lee, Z. (2023). Cyberslacking in the workplace: Antecedents and effects on job performance. *MIS Quarterly* Forthcoming. https://misq.umn.edu/cyberslacking-in-the-workplace-antecedents-and-effects-on-job-performance.html

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, *27*(3), 425–478. https://doi.org/10.2307/30036540

Venkatesh, V., Thong, J. Y. L., & Xu, X. (2012). Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. *MIS Quarterly*, *36*(1), 157–178. https://doi.org/10.2307/41410412

Vercruyssen, M., & Hendrick, H. W. (2011). *Behavioral research and analysis: An introduction to statistics within the context of experimental design*. CRC Press.

Viswanathan, M. (2005). *Measurement error and research design*. SAGE Publications.

Wang, C. L., Ahmed, P. K., & Rafiq, M. (2008). Knowledge management orientation: Construct development and empirical validation. *European Journal of Information Systems*, *17*(3), 219–235. https://doi.org/10.1057/ejis.2008.12

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. https://doi.org/10.1016/j.ijresmar.2010.02.004

Willis, G. B. (2004). *Cognitive interviewing: A tool for improving questionnaire design*. Sage Publications.

Wilson, T. D., & Dunn, E. W. (2004). Self-knowledge: Its limits, value, and potential for improvement. *Annual Review of Psychology*, *55*, 493–518. https://doi.org/10.1146/annurev.psych.55.090902.141954

Winer, B. J. (1962). *Statistical principles in experimental design*. McGraw-Hill Humanities.

Winter, S. J., Saunders, C., & Hart, P. (2003). Electronic window dressing: Impression management with websites. *European Journal of Information Systems*, *12*(4), 309–322. https://doi.org/10.1057/palgrave.ejis.3000470

Wixom, B. H., & Todd, P. A. (2005). A theoretical integration of user satisfaction and technology acceptance. *Information Systems Research*, *16*(1), 85–102. https://doi.org/10.1287/isre.1050.0042

Wunderlich, P., Veit, D. J., & Sarker, S. (2019). Adoption of sustainable technologies: A mixed-methods study of German households. *MIS Quarterly*, *43*(2), 673–691. https://doi.org/10.25300/MISQ/2019/12112

Yao, G., Wu, C. H., & Yang, C. T. (2008). Examining the content validity of the WHOQOL-BREF from respondents' perspective by quantitative methods. *Social Indicators Research*, *85*(3), 483–498. https://doi.org/10.1007/s11205-007-9112-8

Yoo, Y. (2010). Computing in everyday life: A call for research on experiential computing. *MIS Quarterly*, *34*(2), 213–231. https://doi.org/10.2307/20721425

Zimmermann, A., Lioliou, E., & Oliveira, J. S. (2020). How do offshoring-related changes in job characteristics affect onshore managers' affective organizational commitment? The moderating role of perceived organizational valence. *Journal of Information Technology*, *35*(4), 316–336. https://doi.org/10.1177/0268396220942237

How to cite this article: Pillet, J.-C., Carillo, K. D., Vitari, C., & Pigni, F. (2023). Improving scale adaptation practices in information systems research: Development and validation of a cognitive validity assessment method. *Information Systems Journal*, 1–48. https://doi.org/10.1111/isj.12428

APPENDIX A

KEY BELIEFS PERTAINING TO SCALE ADAPTATION PRACTICES

No.	Key beliefs	How we challenge the status quo	Evidence	Our responses
1	Scale adaptation is a trivial practice that involves only tangential alterations of item wording (e.g., change of system name)	Substantial modifications of item wording are common, and some are akin in magnitude to creating new measures	A third of adapted scales demonstrate less than 50% semantic similarity with the baseline scales from which they are derived (cf. Appendix B, $n = 54$ scales)	Propose a method that researchers can use to gauge the performance of an adapted scale from a cognitive validity standpoint
2	Cognitive validity requirements are met as long as respondents can smoothly process item content	Examining cognitive validity solely through the lens of readability could obfuscate other important cognitive validity issues	PICs can interfere with the cognitive operations that take place at the comprehension, retrieval, judgement, and response stages (cf. Table 2)	Theorise all the important facets of the cognitive validity concept and differentiate it from the related concept of content validity
3	Current practices are effective in dealing against the proliferation of poorly adapted scales	There are some ambiguities in the way cognitive validity should be assessed, and in how the empirical data resulting from this activity should be interpreted	About 2% of the papers using a self-reported measure examine cognitive validity using a formal test (cf. Table 1)	Making the execution and interpretation of cognitive validation more transparent, consistent, and reliable could contribute to more frequent assessments
4	Problematic items resulting from poor scale adaptations are easy to detect and PICs can effortlessly be diagnosed using existing techniques	Detecting problematic items is challenging as even subtle wording alterations can influence how people process item content	Experts in item review panels tend disagree be because they have different judgements with regards to items' potential problems (cf. Study 2)	Minimise the influence of subjective judgements in cognitive validity assessments at both the data collection and statistical analysis stages
5	The impact of poor adaptations is minimal because adapted scales remain confined to the study using the scale	Poor adaptations diffuse across the field as scholars re-use adapted scales or parts of them	Authors routinely build on each other's scale to create measures (cf. Appendix B; "cascading adaptations" from Heggestad et al., 2019, p. 2614)	Offer IS researchers a way to benchmark the cognitive validity of their scale against venerable IS measures

APPENDIX B

SEMANTIC SIMILARITY OF ADAPTED SCALES WITH THEIR BASELINE

Note: similarity scores were computed using the http://wordvec.colorado.edu/website, "LSA" embedding method, "General reading 1st year college" embedding space. BIF = Belief in Fate scale (Chan et al., 2009): "Many things in life are predetermined"; "Fate determines one's successes and failures"; "Destiny is what determines the course of someone's life."

Baseline scale (source)	Adapted scale (source)	Baseline-BIF similarity	Baseline-adapted similarity
Transaction convenience (Seiders et al., 2007)	Online Transaction Speed Convenience (Trenz et al., 2020)	3%	53%
Post-benefit convenience (Seiders et al., 2007)	Online Post-Transaction Speed Convenience (Trenz et al., 2020)	10%	44%
Post-benefit convenience (Seiders et al., 2007)	Online Post-Transaction Effort Convenience (Trenz et al., 2020)	10%	30%
Outcome Measurability (Kirsch et al., 2002)	Outcome Control (Remus et al., 2020)	5%	50%
Clan Control (Kirsch et al., 2002)	Clan Control (Remus et al., 2020)	7%	78%
Input Control (Snell, 1992)	Input Control (Remus et al., 2020)	14%	37%
Interruption overload (Chen & Karahanna, 2018)	Perceived interruption overload (Tams et al., 2020)	10%	39%
Work climate for family role (Kossek et al., 2001)	Work–Home Boundary Management Support (Benlian, 2020)	14%	64%
Denial (Carver, 1997)	Denial of ambivalence (Qahri-Saremi & Turel, 2020)	12%	41%
Planning (Carver, 1997)	Planning to Cope (Qahri-Saremi & Turel, 2020)	18%	85%
Addiction (Turel & Serenko, 2012)	SNS addiction (Tarafdar et al., 2020)	32%	87%
Social benefits (Caplan, 2002)	Mobile social interaction (Kuem et al., 2020)	6%	90%
Withdrawal (Caplan, 2002)	Withdrawal (Kuem et al., 2020)	8%	79%
Conflict (Charlton, 2002)	Conflict (Kuem et al., 2020)	29%	61%
Portability (Okazaki & Mendez, 2013)	Ubiquity (Mäntymäki et al., 2020)	4%	48%
IS continuance intention (Bhattacherjee, 2001)	Intention to upgrade/retain subscription (Mäntymäki et al., 2020)	3%	53%
New R&D fields and sources (Cassiman et al., 2005)	Inter-firm Knowledge Distance (Cui et al., 2020)	3%	13%
Product effectiveness (Pavlou & El Sawy, 2006)	Product Effectiveness (Cui et al., 2020)	5%	85%
Process Efficiency (Pavlou & El Sawy, 2006)	Process Efficiency (Cui et al., 2020)	6%	67%
		7%	92%

Baseline scale (source)	Adapted scale (source)	Baseline-BIF similarity	Baseline-adapted similarity
Job complexity (Morgeson & Humphrey, 2006)	Perceived increase in job complexity (Zimmermann et al., 2020)		
Organisational valence (Holt et al., 2007)	Perceived organisational valence (Zimmermann et al., 2020)	8%	82%
Cyberchondria (Jokić-Begić et al., 2019)	Cyberchondria (Laato et al., 2020)	4%	62%
Joy (Lowry et al., 2012)	Joy (Silic & Lowry, 2020)	1%	88%
Self-Efficacy (Milne et al., 2002)	Security response efficacy (Silic & Lowry, 2020)	3%	30%
Seller Uncertainty (Dimoka et al., 2012)	Pre-Purchase Seller Uncertainty (Al-Natour et al., 2020)	6%	80%
Seller Uncertainty (Dimoka et al., 2012)	Post-Purchase Seller Uncertainty (Al-Natour et al., 2020)	6%	70%
Product Uncertainty (Dimoka et al., 2012)	Pre-Purchase Product Uncertainty (Al-Natour et al., 2020)	19%	62%
		Mean = 9% StDev = 8%	Mean = 62% StDev = 22%

APPENDIX C

EVIDENCE OF THE UTILITY OF THE PROCEDURE (STUDY 2)

TABLE C1 Instructions *perceived privacy risk* scale (Control condition)

***Study background*: The following items relate to privacy concerns that could arise prior to the installation of Smart Meter Technology (SMT) in someone's household. An SMT is an electronic device that records the consumption of electric energy and communicates the information to the electricity supplier for monitoring and billing.**	
***Conceptual definition*: Perceived privacy risk (PPR) refers to the potential loss of control over personal information, such as when information about a person is used without his or her knowledge or permission. Concerns about privacy risk evoke consumers' scepticism about using an SMT and negatively affect adoption intentions. Privacy risk is also related to consumers' anxiety regarding energy suppliers' abuse of their private consumption data.**	
ID	**PPR items**
Item1	Using SMT could lead to a loss of control over the privacy of my personal data.
Item2	Using SMT could lead to a loss of my privacy, because my energy consumption data could be used without my knowledge.
Item3	My personal data will not be used for any purposes not related to SMT.
Item4	My personal data that is gathered due to the usage of SMT would not be sold to third party providers.
Item5	I am concerned about the data security of SMT.
Item6	Internet hackers might take control of my payment and consumption data if I would use SMT.
Item7	The databases that are used to save my consumption data are protected against unauthorised access.

TABLE C2 Instructions *exploratory task adaptation* scale (Control condition)

***Study background*: The following items are concerned with the use of mobile phones by workers in the course of their daily job. Mobile phones are considered malleable technologies that can be tailored to fit work-related requirements and that can, conversely, alter someone's work structure. These adaptations often occur through trials and experiments and are thus exploratory in nature.**	
***Conceptual definition*: Exploratory task adaptation (ETA) behaviours involve a transition to something that disregards convention in the workplace. These behaviours include the unintended or unexpected application of the technology to new work tasks or processes during the introduction of a technology. Conversely, users may develop and create non-standard ways of interpreting the technology that are qualitatively different than that which previously existed.**	
ID	**ETA items**
Item1	I try hard to figure out how to perform work-related tasks in new places and settings that were not possible without my current mobile phone.
Item2	I strive to find ways to take on new work responsibilities by using my current mobile phone.
Item3	My current mobile phone has allowed me to frequently attempt new tasks I could not do in the past.
Item4	Overall, use of my current mobile phone has enabled me to try new and different work related tasks.
Item5	I have developed a way of using my mobile phone which deviates from the standard usage.
Item6	I have used at least one mobile phone feature or capability in an usual manner which the vendor does not encourage.
Item7	I have modified something on my mobile phone to use it in a non-standard way.

TABLE C3 Instructions *perceived privacy risk* scale (Treatment condition)

	Item mean ratings							
	Comprehension difficulty		Social desirability		Response inducement		Knowledge assumption	
	Ambiguity	Complexity	Moralistic	Egoistic	Suggested	Implied	Practical	Theoretical
PPR1	4.66	3.51	4.63	4.66	5.20	5.03	4.46	4.66
PPR2	4.43	4.26	4.40	4.69	5.20	4.69	4.66	4.86
PPR3	4.00	3.11	4.09	4.09	4.60	4.49	5.11	5.57
PPR4	3.60	3.89	4.11	3.83	4.20	4.57	4.97	5.11
PPR5	4.31	3.03	4.54	4.34	5.06	5.00	4.54	4.77
PPR6	4.34	3.74	4.40	4.54	5.49	5.23	4.80	4.74
PPR7	3.43	4.26	4.14	4.26	4.26	4.74	5.51	5.43
Scale	4.11	3.69	4.33	4.34	4.86	4.82	4.87	5.02

Note: An "item bias" score from 1 to 7 was computed based on this assessment. High scores indicate that the item is likely to be heavily biased; thus, it requires specific attention to the aspect being evaluated. Items may be heavily biased on a given aspect and acceptable on another aspect. Use the information provided in the table below as a guide in the item-revision process. Notes: Scores in black are significantly greater than the mean scale score (p-value <0.05).

TABLE C4 Instructions *exploratory task adaptation* scale (Treatment condition)

	Item mean ratings							
	Comprehension difficulty		Social desirability		Response inducement		Knowledge assumption	
	Ambiguity	Complexity	Moralistic	Egoistic	Suggested	Implied	Practical	Theoretical
ETA1	4.57	4.54	5.09	4.31	4.46	4.23	3.97	4.09
ETA2	4.49	3.49	5.46	5.34	4.49	4.54	4.11	4.40
ETA3	3.97	3.60	4.91	4.71	4.43	4.51	3.86	4.17
ETA4	3.77	3.14	4.97	4.69	4.14	4.51	4.03	4.29
ETA5	4.29	3.63	3.66	4.83	4.17	3.91	4.69	5.17
ETA6	4.66	4.71	3.26	4.51	4.37	4.14	5.14	4.83
ETA7	4.11	3.20	3.34	4.46	3.69	3.86	5.00	5.09
Scale	4.27	3.76	4.38	4.69	4.25	4.25	4.40	4.58

Note: An "item bias" score from 1 to 7 was computed based on this assessment. High scores indicate that the item is likely to be heavily biased; thus, it requires specific attention to the aspect being evaluated. Items may be heavily biased on a given aspect and acceptable on another aspect. Use the information provided in the table below as a guide in the item-revision process. Notes: Scores in black are significantly greater than the mean scale score (p-value <0.05).

APPENDIX D

NORMS FOR EVALUATING COGNITIVE VALIDITY IN IS (STUDY 3)

Towards a dedicated cognitive validity statistic

To facilitate the interpretation of the ratings, the impact of the number of response options used in the procedure has to be neutralised. The *PIC* index (for *problematic item characteristics*) simply divides the average ratings obtained across a scale's item by *a*, the number of response anchors used in the data collection procedure:

$$PIC = \text{average item ratings}/a.$$

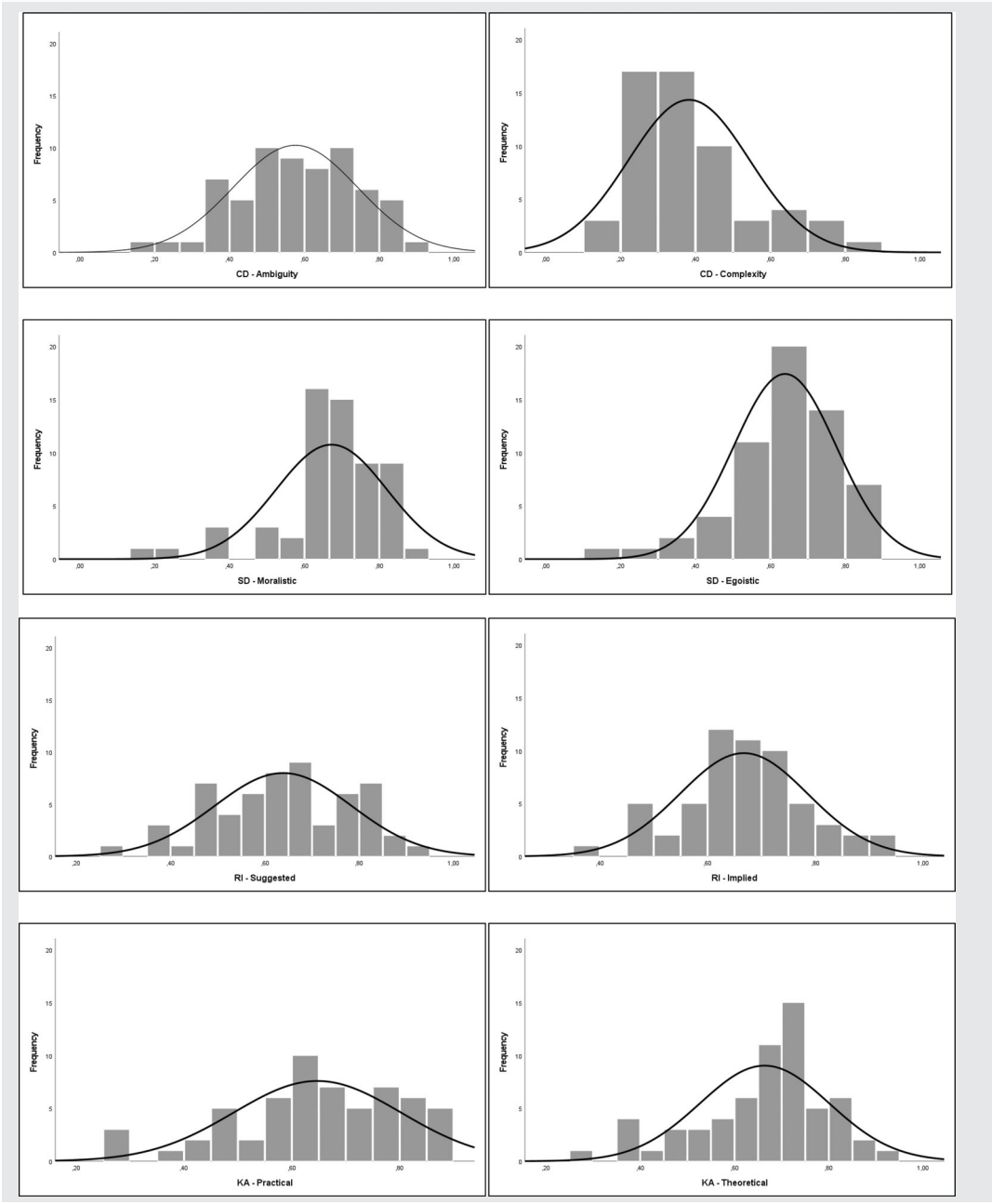
The *PIC* indices take the value of 1 when all judges select the maximum anchor for all the items subjected to the evaluation, which indicates in this hypothetical case that the scale has very poor cognitive properties. Conversely, *PIC* indices near 0 indicate excellent cognitive properties because raters did not detect any problem. Of course, extreme scores nearing either 0 or 1 are very unlikely and a more balanced distribution of ratings between the two bounds is to be expected. The *PIC* index should be used to account for each type of PIC (namely, CD, SD, RI, and

TABLE D1 Mean ratings for the 10 venerable IS scales

	CD		SD		RI		KA	
Scale	Ambiguity	Complexity	Moralistic	Egoistic	Suggested	Implied	Practical	Theoretical
PU	4.00	2.33	5.10	4.91	4.61	4.88	4.74	4.85
	(1.73)	(1.56)	(1.04)	(1.30)	(1.49)	(1.25)	(1.58)	(1.32)
EOU	4.17	2.68	5.33	5.66	4.57	4.73	4.69	4.84
	(1.41)	(1.32)	(1.14)	(1.02)	(1.29)	(1.08)	(1.58)	(1.42)
BI	3.94	2.26	4.64	4.44	3.91	4.11	4.52	4.37
	(1.97)	(1.62)	(1.51)	(1.62)	(1.81)	(1.69)	(1.92)	(1.86)
AFF	3.83	2.27	4.17	3.96	4.53	4.44	4.38	3.90
	(1.35)	(1.34)	(1.04)	(1.01)	(1.22)	(1.12)	(1.39)	(1.86)
FAC	4.05	2.23	4.93	4.52	3.96	4.22	4.95	4.96
	(1.64)	(1.48)	(1.23)	(1.20)	(1.47)	(1.37)	(1.37)	(1.09)
TRST	3.92	2.99	4.32	3.70	4.50	4.93	4.19	4.89
	(1.28)	(1.20)	(1.44)	(1.22)	(1.02)	(0.86)	(1.56)	(1.10)
GIPC	3.99	2.96	4.30	4.06	5.03	4.81	3.43	4.44
	(1.31)	(1.13)	(1.24)	(1.21)	(1.07)	(1.09)	(1.50)	(1.29)
SAT	4.00	2.43	4.92	4.48	4.99	5.10	4.86	4.62
	(1.60)	(1.55)	(1.38)	(1.25)	(1.16)	(1.18)	(1.63)	(1.40)
SEF	4.36	3.79	4.58	4.70	4.41	4.41	4.45	4.62
	(1.18)	(1.21)	(1.16)	(1.07)	(1.51)	(1.18)	(1.26)	(0.98)
OI	4.05	2.98	4.76	4.30	4.21	4.90	5.38	5.28
	(1.33)	(1.14)	(1.47)	(1.42)	(1.41)	(1.13)	(1.14)	(1.24)

Abbreviations: AFF, affect; BI, intention to use; EOU, perceived ease of use; FAC, facilitating conditions; GIPC, global information privacy concern; OI, organisational impact; PU, perceived usefulness; SAT, satisfaction with IT; SEF, computer self-efficacy; TRST, trusting beliefs.

TABLE D2 Distribution of the PIC indices



KA). That is, there is no overarching *PIC* statistic that would aggregate the mean ratings across all the evaluated types of PICs, but rather a collection of *PIC* indices, each one corresponding to a type of problem, that together shed light on various aspects of a scale's cognitive properties.

Definition of criteria for interpreting the cognitive validity statistic

We first sought to identify venerable constructs whose definitions are well understood and used frequently in the literature. The decision to focus on venerable IS scales that have been heavily cited and have demonstrated sound psychometric properties is motivated by the need to set gold standards against which researchers can benchmark the performance of their own scale from a cognitive standpoint. We started with the list of the 10 most common IS constructs following the taxonomy provided by (Larsen & Bong, 2016). The taxonomy provides an overview of the constructs that comprise of a minimum of three items during the period from 1983 to 2009 that appear in *MIS Quarterly* and *Information Systems Research*. The most 10 common constructs clusters are *Individual-level usefulness*, *Ease of use*, *Intention to use*, *Affect towards technology use*, *Facilitating conditions*, *Individual-level trust*: *Trustworthiness*, *General concerns about information privacy*, *Satisfaction with technology*, *Technology self-efficacy*, *Organisation-level usefulness*.

We then searched for the most prominent scales within each construct cluster. Because the same construct might be measured using different scales or different versions of the same scale, we initially shortlisted more than one scale per construct cluster. We used the INN website[7] (Larsen & Bong, 2016) to supplement the scale identification process. We also relied on literature reviews of the construct as those would typically point at commonly used scale. For example, the identification of scales to measure privacy-related concerns in IS was informed by literature reviews on the topic (Bélanger & Crossler, 2011; Li, 2012; Pavlou, 2011). We narrowed down the set of scales to retain for each cluster to scale originating from papers that have received more than 1000 citations.[8] The list of 10 venerable IS scales is reported on Table 9 of the main manuscript. Subjecting these scales to the method resulted in the mean ratings provided in Appendix B (on a scale from 1 to 7).

[7] https://inn.theorizeit.org/

[8] According to Google Scholar, accessed in June 2021.

APPENDIX E

ILLUSTRATED IMPLEMENTATION STEPS

TABLE E1 Recommended implementation steps

Step	Implementation considerations and recommendations
1. Defining the Research Objectives	*Purpose*: does the method fulfil a need to collect useful information that will serve to improve the scale items (i.e., developmental purpose) or will the method primarily serve to extract evidence of the validity of the measures (i.e., instrumental purpose)? *Relation with other item assessment methods*: have or will other techniques to detect problematic items be used? What method will be implemented to assess content validity, and how will the procedures complement one another? *Origin of the items*: what is the maturity of the scale items to assess? Are the items completely new or have they been adapted from another study? Have the items been translated for the purpose of the project?
2. Identifying the Items to Assess	We recommend using at least 5 items so that meaningful statistical tests of the mean difference can be performed. Including items from conceptually related scales for comparison purposes would allow a more refined assessment.
3. Implementing the Rating Procedure	*Selection of the PIC measures*: using all eight measures will provide a comprehensive assessment of the scale, but some measures may not be relevant (e.g., KA) for some research instruments. *Design of the task*: researchers should implement procedures to prevent rater fatigue, randomise the appearance of blocks and items to prevent response-order effects, and include attention-checks. We recommend using fully labelled response options in a Likert-type format. *Rater number and profile*: the number of raters should be within the range of 30 to 50 to ensure that the responses are normally distributed. We recommend recruiting raters who can perform cognitively demanding tasks and who have a good command of the language in which the items are written (e.g., undergraduate students), unless item specificities demand otherwise.
4. Diagnosing Potential Problems	*At the level of the individual items*: detect problematic items by comparing their ratings to the combined ratings for the rest of the items on each PIC category. We recommend using a repeated one-way ANOVA planned contrast. *At the level of the scale*: examine the *PIC* indices for each category to diagnose systematic issues with the scale. We recommend benchmarking those results with those of venerable IS scales (see Study 3). *PIC* indices within the 0.70 to 1.00 range would signal a potential problem with the scale, except from *Complexity*, where any score above 0.50 would demand careful examination of the item.
5. Implementing the Appropriate Remedies	*Item-specific remedies*: items should be revised regarding the problem(s) that has (have) been diagnosed in the previous step. *Scale-specific remedies*: researchers should consider whether the problem can be attenuated by redesigning the scale's items or if specific procedural and statistical remedies should be implemented.

TABLE E2 Overarching instructions

CONTEXT:

Statements sometimes contain subtle differences that can influence how people respond to them. Both the structure of a statement and the choice of words can have a considerable influence in how people respond.

TASK:

Your task consists of evaluating the quality of a list of statements that are typically used in technology-related research. You will successively evaluate the quality of the statements on the following four unrelated aspects:

- Is the intended meaning of the statement difficult to grasp?
- Is the statement prone to over- or underreporting thoughts or behaviours?
- Is the statement subtly encouraging respondents to endorse it?
- Is the statement difficult to process for people who are not experts on the survey topic?

Note that the reasoning for this task differs from what you are used to. Make sure that you do not respond to the content of the statements but evaluate them by using the header questions.

Here is a small rehearsal for you to practice.

The statement contains emotionally loaded terms:

	1 Strongly disagree	2 Disagree	3 Somewhat disagree	4 Neither agree nor disagree	5 Somewhat agree	6 Agree	7 Strongly agree
All things considered, this technology could raise confidentiality issues.	○	○	○	○	○	○	○
All things considered, this technology could violate my privacy.	○	○	○	○	○	○	○

In this example, the term "violate privacy" is probably more emotionally charged than the term "raise confidentiality issues". Such a subtle difference can influence responses on a large scale.

You are now ready to start the task. Good luck!

TABLE E3 Block-specific instructions

COMPREHENSION DIFFICULTY

Respondents may find it difficult to intuitively comprehend a statement because

- It has multiple meanings, or
- It uses a complex grammatical structure.

For example, respondents are unlikely to understand the meaning of the statement "I support civil rights because discrimination is a crime against god" {1—strongly disagree to 5—strongly agree} because it is unclear whether they are supposed to report their support on civil rights or the extent to which they think that discrimination is an affront to a deity.

You will now assess the extent to which a statement is difficult to comprehend intuitively.

SOCIAL DESIRABILITY

A statement may be worded in a way that prompts respondents to provide a biased response because it points to attributes that are viewed by others (friends, family, colleagues, etc.) as desirable to possess.

For example, respondents are more likely to overestimate their actual behaviour in the statement"I put in extra time to help coworkers with work-related problems" {1—strongly disagree to 5—strongly agree} because assisting others is a socially desirable behaviour.

You will now assess the extent to which a statement is prone to socially desirable responding.

RESPONSE INDUCEMENT

Statements implicitly encourage respondents to provide a certain response because

- The personal preference of its author is apparent in its wording, or
- It logically implies a certain response.

For example, the statement "Family life often suffers because men concentrate too much on their work" {1—strongly disagree to 5—strongly agree} leads respondents to endorse the statement because the word "suffers" is a loaded word. It also presupposes that all men concentrate too much on their work, which hints at the personal opinion of the author of the statement.

You will now assess the extent to which a statement induces respondents to respond in a certain direction.

KNOWLEDGE ASSUMPTION

Researchers can overestimate the extent to which respondents are knowledgeable about the topic of a survey. When this phenomenon occurs, respondents cannot provide sensible responses. In technology-related research, this situation occurs when researchers assume that respondents:

- Have extensive hands-on experience using the technology, or
- Are well informed about IT-related issues.

For example, to provide a reasonable response to the statement "My organisation shares information when it is legally bound to do so" {1—strongly disagree to 5—strongly agree}, respondents should have an understanding of which information an organisation is legally bound to communicate, which is unlikely unless they are in a very specific position (e.g., corporate lawyer).

You will now assess the extent to which a statement assumes that respondents have thorough knowledge in the information technology (IT) domain.

TABLE E4 Sample visualisation of the rating task

The statement has several possible meanings:

	1 Strongly disagree	2 Disagree	3 Somewhat disagree	4 Neither agree nor disagree	5 Somewhat agree	6 Agree	7 Strongly agree
I have modified something on my mobile phone to use it in a nonstandard way.	○	○	○	○	○	○	○
Overall, use of my current mobile phone has enabled me to try new and different work related tasks.	○	○	○	○	○	○	○
I have used at least one mobile phone feature or capability in an usual manner which the vendor does not encourage.	○	○	○	○	○	○	○
Please select "1-Strongly disagree" to prove that you are paying attention.	○	○	○	○	○	○	○
I strive to find ways to take on new work responsibilities by using my current mobile phone.	○	○	○	○	○	○	○
I have developed a way of using my mobile phone which deviates from the standard usage.	○	○	○	○	○	○	○
I try hard to figure out how to perform work-related tasks in new places and settings that were not possible without my current mobile phone.	○	○	○	○	○	○	○

TABLE E5 Example reporting template

	Comprehension difficulty		Social desirability		Response inducement		Knowledge assumption	
	Ambiguity	Complexity	Moralistic	Egoistic	Suggested	Implied	Practical	Theoretical
Item 1	4.66	3.51	4.63	4.66	5.20	5.03	4.46	4.66
Item 2	4.43	4.26	4.40	4.69	5.20	4.69	4.66	4.86
Item 3	4.00	3.11	4.09	4.09	4.60	4.49	5.11	5.57
Item 4	3.60	3.89	4.11	3.83	4.20	4.57	4.97	5.11
Item 5	4.31	3.03	4.54	4.34	5.06	5.00	4.54	4.77
Item 6	4.34	3.74	4.40	4.54	5.49	5.23	4.80	4.74
Item 7	3.43	4.26	4.14	4.26	4.26	4.74	5.51	5.43
Average	*4.11*	*3.69*	*4.33*	*4.34*	*4.86*	*4.82*	*4.87*	*5.02*
pic(*)	0.59	0.53	0.62	0.62	0.69	0.69	0.70	0.71

Note: Scores in black are significantly greater than the mean scale score (ANOVA contrasts, $p < 0.05$). (*) *pic* = average item ratings/7 (number of anchors).