Exploring the Impact of Randomized Item Selection on Content Sampling Error in Psychological Measurement

Berre Deltomme, Bert Weijters

Department of Work, Organization and Society

Faculty of Psychology and Educational Sciences

Ghent University

Henri Dunantlaan 2

9000 Gent

Belgium

Corresponding author: Berre.Deltomme@ugent.be

Bert.Weijters@ugent.be

Abstract

Psychological research heavily relies on the use of multi-item self-report scales. Traditionally, each participant is administered the same set of items, under the assumption that these items comprehensively represent the intended construct. However, an inherent challenge arises due to content sampling error, stemming from the disparity between the utilized subset of items and the complete universe of possible items. In this study, we explore whether randomly selecting items per respondent from a validated item pool might counter content sampling error on an aggregate level. We compare the application of random item selection to traditional survey methods. Using the construct of 'Pro-Environmental Behavior' (PEB) as an example, respondents were randomly assigned to either the traditional or randomized approach. For the randomized approach, one item was randomly selected for each of the ten PEB domains from a pool of ten possible items for each respondent. Four scales, separated by a filler task were administered. For the traditional approach, two fixed scales were created with one item per domain and administered two times, again separated by a filler task. By correlating the outcomes we can assess the convergent validity. The traditional approach shows higher correlations between scales, but variance decomposition reveals that the randomized condition captures a broader range of content. Lower correlations in the randomized condition are due to higher item and residual variance. The potential benefits of using validated item pools with random item sampling are discussed, with a focus on both psychometric improvements and researcher involvement in the measurement process.

Keywords: Psychological measurement, self-report scales, content validity, content sampling error, psychometrics

Public significance statement

Psychological measurement presents enduring complexities, with challenges persisting over decades. In this paper, we introduce and explore a novel approach to scale-based psychological measurement. Our

focus lies in investigating whether administering a randomized selection of items (in contrast to a fixed set of items) to each respondent within a study could potentially bolster content validity. Our findings suggest that this method may hold promise as an effective avenue for enhancing measurement validity.

Introduction

In contemporary psychological research, the assessment of psychological constructs often relies on self-report scales because of their ease of implementation and analysis. The availability of numerous validated scales has further promoted their widespread use. However, recent developments in the field have highlighted significant challenges in the measurement of psychological constructs. Studies across diverse subdisciplines of psychology, including environmental and clinical psychology as well as organizational and emotion research, have uncovered important problems. For instance, Antonakis and colleagues (2016) critically reviewed the conceptualization and measurement of charisma in sociological and organizational sciences, identifying major challenges such as imprecise definitions, confusion with other constructs, reliance on inadequate questionnaire measures, and improperly specified causal models. Similarly, Fried (2017) criticized the heterogeneity within seven well-known depression scales and showed little overlap between the symptoms depicted in the different scales. Weidman and colleagues (2017) shed light on problems in emotion assessment, emphasizing imprecise measurement and conceptual ambiguity. Warnell and Redcay (2019) explored Theory of Mind, the ability to understand others by attributing mental states to them, in both childhood and adulthood, revealing minimal coherence across different measures. Deltomme et al. (2023) demonstrated the lack of strong correlations among established self-report scales measuring 'Pro-Environmental Behavior', rendering them essentially non-interchangeable.

The issue was also highlighted in a recent editorial article in the section of quantitative psychology and measurement in Frontiers in Psychology with the self-explanatory title ‘Persistence of measurement problems in psychological research’ (Meier, 2023) in which the author addresses the necessity of exploring new avenues in psychological measurement to counter the above described long-standing problems.

This prevalent concern in psychological research can be dissected into two crucial components: researcher involvement and psychometric factors. The researcher involvement aspect pertains to the lack of diligence often exhibited by researchers when approaching measurements. Flake & Fried (2020) coined this phenomenon as a 'measurement schmeasurement attitude', highlighting the tendency of researchers to employ questionable measurement practices.

The field of psychometrics considers the tools and measurement theories that can be applied in the measurement process and relies on foundational work by pioneers such as Cronbach (1949), Thurstone (1927), and Nunnally (1967). These seminal works provide a cornerstone upon which the field of psychometrics can continue to evolve, integrating contemporary insights and methodologies to cultivate a more nuanced and comprehensive understanding of psychological measurement. Even incremental enhancements in measurement methodologies can yield significant and impactful advances (Meier, 1994, 2008). However, most progress seems to have focused on statistical testing, exemplified by the rise of structural equation modeling and network modeling, while the crucial measurement phase has received less attention than it deserves (Lilienfeld & Strother, 2020; Fried & Flake, 2018).

Despite advancements in statistical procedures, the fundamental principle of 'garbage in, garbage out' remains unchanged when dealing with poorly measured constructs. Longstanding problems with the validity of self-report scales or method effects remain an important problem in psychological measurement. Furthermore, Baumgartner et al. (2021) point out that survey researchers commonly underestimate the potential biasing effects of common method variance. Method effects refer to the observation that every quantitative outcome of a measure at least partially reflects the specific methodology that has been employed. This has a nonnegligible influence on the variance in self-report measures. For instance, Cote and Buckley (1987) investigated 70 datasets and found that, on average, traits account for less than 50% of the variance in construct measures, while method variance accounts on average for 26.3% and random error variance for 32% of the total variance.

Measurement theories

A useful way to conceptualize the relationship between observations (here, items) and the constructs they intend to measure is offered by Behavior Domain Theory (McDonald 2003). Behavior Domain Theory (BDT) focuses on categorizing behaviors into domains based on functional characteristics. Validity in Behavior Domain Theory is established by defining the identity of behavior domains and ensuring adequate sampling from these domains. BDT conceptualizes the relation between a domain and its indicators as a statistical (sampling) one, where indicators need to be sampled from a homogeneous common domain in an effort to obtain a representative sample. The homogeneity of these domains can be, but need not be, the consequence of a common cause (Markus & Borsboom, 2013). For BDT, construct validity, which refers to the degree to which a measurement instrument correctly measures the theoretical construct it aims to measure (Borsboom et al., 2004), is contingent upon how well the scale items represent the items in the theoretical universe corresponding to that construct. However, achieving a valid representation of this universe is challenging, leading to what is known as content sampling error. In the context of BDT in psychometrics, content sampling error refers to the error that occurs when the items selected to measure a construct do not adequately represent the entire domain of behaviors, skills, or attributes that define that construct. This type of error is a significant concern when developing or evaluating psychological measures, as it directly impacts their validity.

Latent constructs are abstract concepts that cannot be directly observed or measured, necessitating reliance on observable indicators, commonly in the form of self-report items in psychology. The way items need to be selected and implemented as indicators of a latent construct depends on whether the relation between the construct and the indicators is best conceptualized as reflective or formative. As explained by Jarvis et al. (2003), the direction of causality is from construct to measure in reflective models, but from measure to construct in formative models. In the reflective model, measures are therefore expected to be correlated and dropping an indicator from the measurement model does not

alter the meaning of the construct (Baumgartner & Weijters, 2019; Edwards, 2010; Howell et al., 2007). By contrast, formative models do not imply that measures are correlated, and dropping an indicator may alter the meaning of the construct. To illustrate, socioeconomic status is a formative construct that is typically thought of as a composite of educational level, income, and occupation (Jarvis et al. 2003).

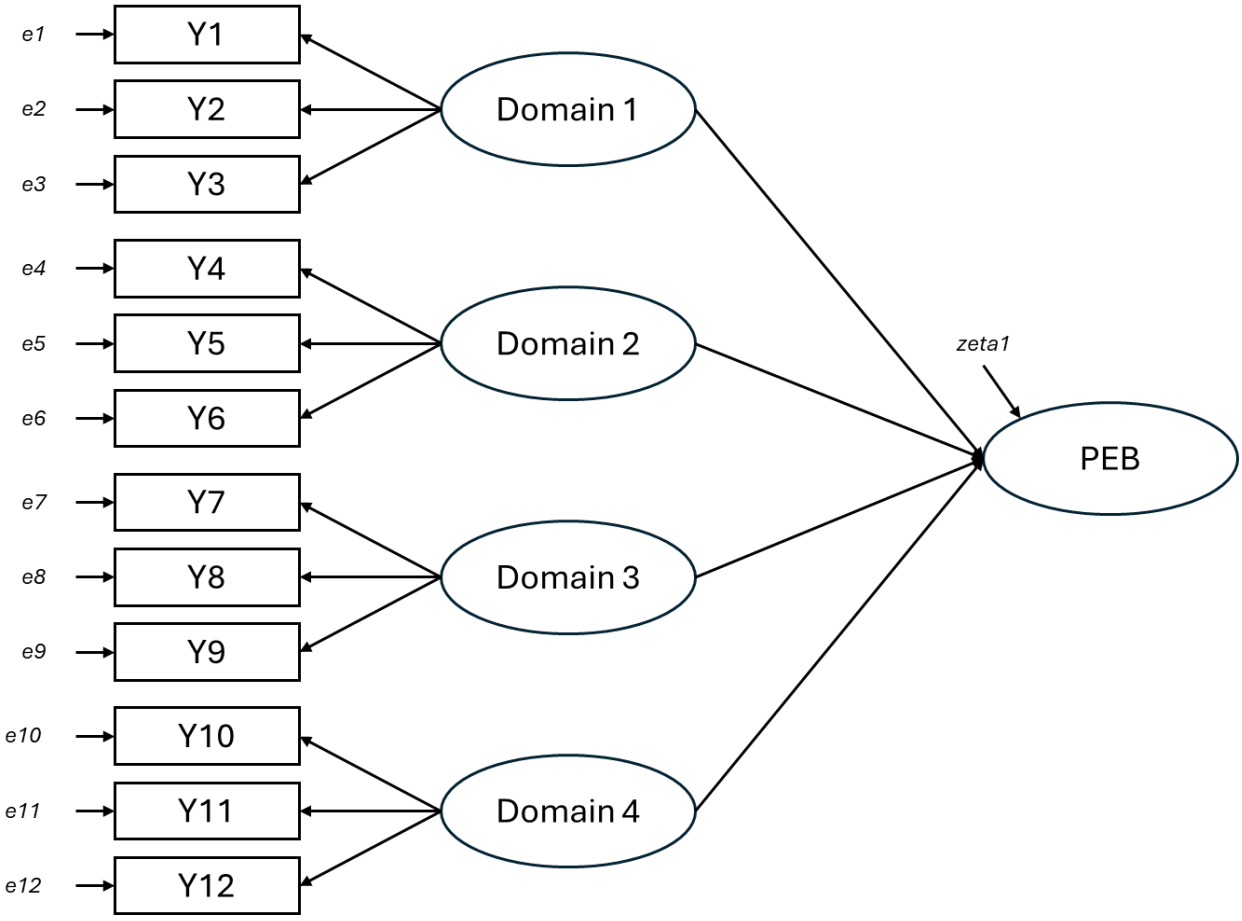
According to Jarvis et al. (2003), “a construct should be modeled as having formative indicators if the following conditions prevail: (a) the indicators are viewed as defining characteristics of the construct, (b) changes in the indicators are expected to cause changes in the construct, (c) changes in the construct are not expected to cause changes in the indicators, (d) the indicators do not necessarily share a common theme, (e) eliminating an indicator may alter the conceptual domain of the construct, (f) a change in the value of one of the indicators is not necessarily expected to be associated with a change in all of the other indicators, and (g) the indicators are not expected to have the same antecedents and consequences. On the other hand, a construct should be modeled as having reflective indicators if the opposite is true...” To decide which of the two models to use, researchers can run thought experiments using these questions as a guideline. The decision on how to model latent variables is important, as it has been shown to affect substantive conclusions (Rhemtulla et al. 2020).

Constructs can integrate both formative and reflective measurements at different levels of the measurement model (see Figure 2 in Jarvis et al., 2003). An example of this is the construct of Pro-Environmental Behavior (PEB) as conceptualized in this manuscript. PEB measures the frequency at which an individual engages in behaviors that positively impact the environment. Since the items used to assess PEB may not share the same antecedent and are thus not expected to necessarily correlate, the construct is not inherently reflective. However, within specific domains, items may be correlated, indicating a degree of interchangeability suggesting that they reflect similar underlying aspects within that domain. For instance, the 'energy' domain includes indicators related to an individual's energy use. It is likely that a person who practices energy-efficient behavior (driven by an underlying motivator such as cost-saving or

environmental concern) will score highly on energy-related indicators, as their energy-conscious behavior manifests across various energy-related activities. Each domain may thus be influenced by a distinct underlying motivator, leading to homogeneity within the domain due to a common factor. The score for each domain then serves as a formative indicator for an overall PEB frequency score. While items within a domain may be interchangeable and drawn from a larger set, maintaining the domain structure is crucial to preserving the formative nature of the PEB composite score. As such, the relationship between a domain and its indicators links to BDT, while the relationship between the domains and the PEB fits the conditions put forward by Jarvis et al. (2003) for formative measurement. For example, omitting a domain that measures recycling behavior would impact the overall PEB score. However, the reverse is not necessarily true: an increase of one standard deviation in the PEB score does not imply that the score for each domain would also increase by one standard deviation. This is because each of the domains might be influenced by another motivator. For example, someone might be energy efficient out of monetary reasons, often ride the bike because of enjoyment and eat organic food from a health motive. As such, PEB is not the underlying latent factor causing changes in the item responses. Figure 1 represents a path model of our conceptualization of PEB, which corresponds to a Type II measurement model in the typology proposed by Jarvis et al. (2003). Note that Figure 1 is a stylized depiction of the PEB model, in most applications the number of domains/items might be higher.

Given the relationship between reflective indicators and their construct, we suggest that reflective indicators can in principle be randomly sampled (independently for each individual respondent); by contrast, the different domains of PEB cannot be randomly sampled but ideally need to all be represented in a PEB measure (to the extent that this is feasible).

Figure 1. Pro-Environmental Behavior model



Note: The relationship between domains and items is reflective. The relationship between the domains and the PEB composite score is formative. Note that this is merely an example, and PEB may consist of additional domains.

To measure psychological constructs via self-report scales, researchers typically employ a set of predetermined items developed through a systematic scale construction process. The process of scale construction involves researchers defining the construct's domain, creating items that thoroughly cover that domain, and subsequently refining the scale to produce a valid and reliable measure (Churchill, 1979, p. 70). Construct validity is achieved when the measure genuinely assesses what it intends to measure (Churchill, 1979; Cook & Buckley, 1979). Thus, it is important to take a representative subset of all possible indicators so that the entire construct domain is covered. The more representative this subset is, the less

content sampling error will be present. However, it is theoretically impossible to correctly sample from the full theoretical item universe. Therefore scale construction commonly involves initially generating a large item pool, which is then seen as a proxy of the universe. Still, a sample will at best offer an approximation of the full universe it is taken from (without ever completely representing it). Through subsequent steps, this pool is then refined until a set of items remains with proven reliability and content validity.

Content sampling error

The discrepancy between the sample of items used in a scale and the ideal representative sample is referred to as content sampling error. Several theories have already tried to deal with measurement error such as Total Survey Error (Groves et al., 2009), Multiple Matrix Sampling (Shoemaker, 1973), and Generalizability Theory (Cronbach et al., 1972). The current study most closely aligns with the latter theory. In the current paper, we aim to explore whether randomly sampling items from a validated item pool might help counter the issue of systematic content sampling error in traditional scale building. Secondly, we discuss some potential advantages of using validated item pools in how they may enhance researcher's involvement in the measurement process.

Specifically, we propose to randomly select items per domain for each respondent, while keeping the domains constant for every respondent. This method of random item sampling for each respondent, combined with the use of validated item pools, offers several advantages for both psychometrics and researchers' involvement in the measurement process.

First, it may account for content sampling error and, as such, make the sampling from the possible universe of items a concrete part of the error. The use of validated item pools comes closer to the universe of items since a broader sample of the universe is attained, thus reducing systematic content sampling error. Furthermore, the random selection of items for each respondent introduces a level of randomness to content sampling error, contrasting with the more systematic sampling error found in scales with fixed

items. As the error is random, it is expected that, on an aggregate level, content sampling error will average out (within the used item pool). Note that because of this, the approach only applies to aggregate data (e.g., comparing population means) and not to the outcome of individual psychological tests (e.g., comparing measurement scores of two individual respondents).

Second, using validated item pools with random sampling may help enhance comparability across studies worldwide. Now, researchers use a plethora of fixed self-report scales to measure the same construct. However, these fixed-item scales often lack content overlap, which makes it unclear to what extent they represent the same content domain. For example, a recent study on validated questionnaires for early childhood adversity, which measures the impact of aversive events on child development, showed heterogeneity in the content and considerable differences in the structural properties, having strong implications for the comparability of the scales (Koppold et al., 2023). In contrast, by employing random item sampling for each respondent within a study—and across different studies—the content sampling error becomes random rather than systematic. This randomness distributes the error across respondents and studies, allowing it to cancel out in the larger research field. This enhances the integrity of individual studies and strengthens meta-analytic findings, advancing knowledge in the field.

Third, the use of validated item pools requires researchers to be more actively involved in the validation process, as they may need to omit variables that are not valid for the specific context or respondent sample, prompting critical reflection. For example, items depicting air-conditioning use may be invalid in regions where air conditioners are not commonly used. Or items depicting driving behavior are invalid in a sample that includes youth as they are not yet permitted to drive. As discussed earlier, this may make the researcher more actively involved in the measurement process as researchers may not be able to ‘hide’ behind validated or established self-report scales. This may remedy the often ‘ritualistic’ way in which self-report scales are applied in contemporary research (Lilienfeld & Strother, 2020). While the primary emphasis is on the random sampling of items from the item pool for each respondent in this study,

researchers may also employ purposive sampling from the item pool (Campbell et al., 2020). Purposive sampling, typically utilized for selecting respondents, can also be applied to item sampling. This approach involves choosing items strategically to gather the most relevant and useful information. A commonly used strategy within purposive sampling is stratified sampling, that is, selecting specific items for each domain that provide the most information for the study's objectives or are valid for the context and sample under study. As such, the development of validated item pools might also be useful for purposive sampling.

Fourth, validated item pools can be updated relatively effortlessly, allowing the removal of items that have become outdated due to changing circumstances (e.g., 'I recycle newspapers' might be less relevant in countries where papers are predominantly read on devices). This adjustment can be made without the need to undergo the entire scale validation process anew. Similarly, new and significant behaviors can be seamlessly added to the item pool following a short item validation process. This mirrors the dynamic nature of the world, accommodating constant changes in behaviors driven by technological advancements or evolving knowledge. Hence, self-report scales may serve as dynamic measurement tools, in contrast to the prevailing static usage of self-report scales with fixed items in contemporary practices.

Current study

In light of the identified challenges in psychological measurement, this study seeks to contribute to the field by offering an initial exploration of the efficacy of a random item sampling method compared to the traditional use of self-report scales. We will use the construct of Pro-Environmental behavior (PEB) as an example. We will approach the construct from behavior domain theory and categorize items by similarity features of behaviors in similar domains. For instance, transportation, energy use, and recycling. By adopting a novel approach that incorporates a validated item pool, we aim to address the pervasive issues related to content sampling error, construct validity, and the comprehensive measurement of

multidomain constructs. We will use a composite score for PEB as this most closely aligns with the way we conceptualize it here (Rhemtulla et al. 2020).

Methods

Participants

A convenience sampling method was used. N = 250 respondents filled out the survey (Male = 94, Mage = 42.6). Respondents were recruited through Prolific and received seventy-one pence (0.71£) for their participation. None of the respondents failed the attention check and none indicated that their data were unreliable and should not be used for the analysis.

Procedure

Respondents were randomly assigned to either the traditional or random sampling condition. In the traditional condition, two scales were presented twice, separated by a filler task. Each scale consisted of a fixed set of ten items, one per domain. The order of scales and items within each scale was randomized. In the random sampling condition, respondents followed the same procedure, except that for each of the four scales, items were randomly selected from ten possible options per domain. The order of selected items within each domain was also randomized. The filler task was included to obtain a more valid test-retest correlation for the traditional condition. Before starting, respondents provided informed consent. Figure 2 outlines the procedure.

Figure 2. Procedure

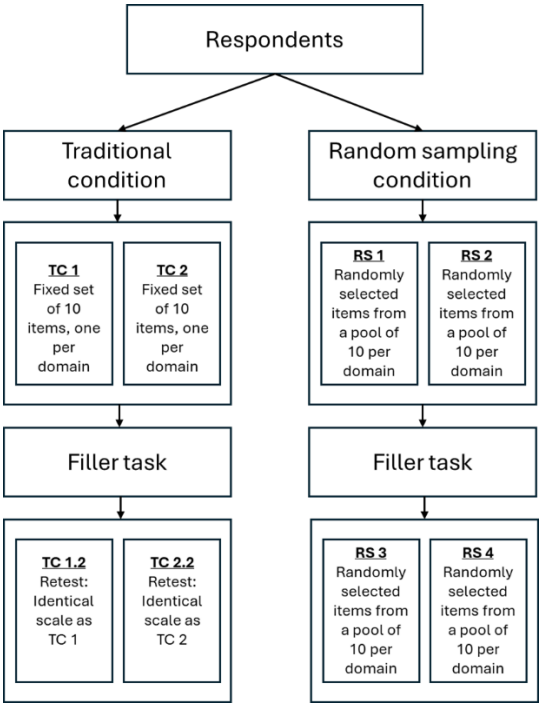
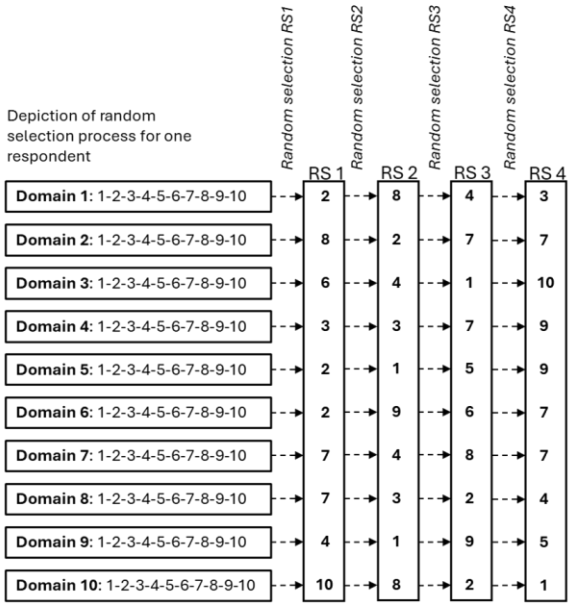


Figure 3. Random selection process



Note: Each column depicts one of the four scales with random item selection out of ten possible items per domain. The items for each domain represent the validated item pools.

Materials

Pro-Environmental Behavior. The item pool used to measure Pro-Environmental Behavior (PEB) was derived from a review of self-report scales for PEB. In this review, 94 papers from the Journal of Environmental Psychology were analyzed to determine if they used self-report scales to measure PEB. This process yielded a list of 404 items. To avoid non-random overlap in content, items depicting similar behaviors (with sometimes different wording) were collapsed by three researchers, reducing the initial 404 items to 215. These 215 items were then reformulated into a sentence starting with 'I'. Finally, the items were categorized into several domains, ten of which will be used in this study. The categorization

was reached by consensus among three researchers. For this study, we selected ten items per domain out of which random items could be drawn (see below). The domains included in this study are: Reducing Waste, Avoiding Harmful Products, Littering, Water Use, Food Consumption, Transportation, Public Sphere Behavior, Recycling, Reuse, and Energy (we acknowledge that this categorization is not conclusive, as it may not cover all facets of PEB and other classifications are possible).

The two scales with fixed items (for the traditional condition) were created by once randomly selecting one item per domain. The items that were used in the fixed item scales were not included in the domains item lists of the randomized condition, however, some items had very similar phrasing to the items included in the pools for the random condition. Responses had to be given on a five-point agreement scale (Totally disagree; Slightly Disagree; Neither agree, no disagree; Slightly agree; Totally agree). The specific items can be found in the Supplement A.

Filler task. The filler task consisted of eight items. For four items, respondents had to solve simple addition and subtraction equations and select the correct option out of three. The other four items consisted of a word task in which respondents had to indicate the antonym of a depicted word (e.g. large-small). The specific filler task items can be found in Supplement A.

Attention check. An attention check was included in the filler task, asking respondents to select the option 'chicken' from three choices. The specific item is provided in Supplement A.

Results

Data analytic plan. Two items were reverse scored: item 9 in TC2 and TC2.1 (identical to TC2): "I prefer to litter when outside than taking my garbage home." The second item that was reverse scored is part of the item pool for the randomized condition, being number 3 under the recycling category: "I put dead batteries in the garbage". After reverse coding, mean scores were calculated for each scale.

By correlating the sum scores of the administered scales, we can evaluate the convergent validity of the scales. Convergent validity gauges the degree to which two measures, intended to assess the same construct, yield similar results. Thus, by correlating the different scales we can investigate to what extent they indeed capture the same construct, which is then indicative of construct validity. For the traditional condition, we investigate the test-retest correlation for the same scales and the cross-correlations (and thus convergent validity) of the different scales with fixed items. In the results, we denote TC1, TC2, and RS1, RS2 the scales that were assessed before the filler task. TC1.2, TC2.2, RS3, and RS4 are the scales that were administered after the filler task. As such for the traditional condition, TC1 is the same scale as TC1.2, and TC2 is the same scale as TC2.2. 'TC' refers to the traditional condition, while 'RS' refers to the random sampling condition.

To understand the influence of the different variance components we conduct a variance decomposition analysis in the second step, which helps us interpret the data from a generalizability theory perspective. In contrast to Classical Test Theory, which divides variance into true and error variance, generalizability theory (abbreviated as G-theory, Cronbach et al. (1972)) accounts for several sources of variance by decomposing the variance according to different facets. In this case, the facets of interest are respondents, items, scales, and time (after versus before the filler task). To calculate the variance components, we use the default Minimum Norm Quadratic Unbiased Estimation method in IBM SPSS Statistics Version 28.0.1.0. In the variance decomposition, time and domain are specified as fixed factors, the other factors are specified as random. The decomposition of the scales is thus reported as follows: *Var(score) = Var(Respondent) + Var(Scale) + Var(Item) + Var(Respondent*Time) + Var(Respondent*Scale) + Var(Respondent*Domain) + Residual Variance*

Analysis results. The means (and standard deviations) are TC1 = 3.88 (.59), TC2 = 3.69 (.49), TC1.2 = 3.85 (.60), and TC2.2 = 3.71 (.49) for the four measures in the traditional condition, and RS1 = 3.37 (.56),

RS2 = 3.35 (.60), RS3 = 3.34 (.60), and RS4 = 3.40 (.60) in the randomized condition. Table 1 reports the correlations.

Table 1. Correlation tables

Traditional		TC1	TC2	TC2.1
	TC2	0.79		
	TC1.2	0.96	0.78	
	TC2.2	0.79	0.95	0.78

Randomized		RS1	RS2	RS3
	RS2	0.57		
	RS3	0.57	0.58	
	RS4	0.49	0.59	0.58

As can be seen from Table 1, the correlations for the random sampling condition are lower compared to the traditional condition. One reason is that TC1 and TC1.2 were identical scales taken at two different moments, as are TC2 and TC2.2 (thus, they reflected test-retest correlations), in the randomized condition this is not the case, and RS1 did not have the same items as RS3 (nor did RS2 have the same items as RS4). The raw correlations clearly bore this out. The test-retest correlation for TC1 and TC1.2 is r=.96 and the correlation for TC2 and TC2.2 is r=0.95. By contrast, the cross-correlations lay between 0.78 and 0.79 (M=0.785). The correlations of the randomly developed scales range from 0.49 to 0.59 (M=0.56), which is still lower than the cross-correlations in the traditional condition. In the second step, we investigate the variance decomposition which is reported in Table 2.

Table 2. Variance decomposition

	Raw estimates		Proportions	
Component	Traditional	Random sampling	Traditional	Random sampling
Var(Resp)	0.188	0.179	14.9%	8.6%
Var(Scale)	0.011	0	0.9%	0.0%
Var(Item)	0.038	0.685	3.0%	32.8%
Var(Resp * Time)	-0.018 [a]	0.007	-1.4% [a]	0.3%
Var(Resp * Scale)	0.036	0.002	2.9%	0.1%
Var(Resp * Domain)	0.509	0.287	40.3%	13.7%
Residual variance	0.499	0.931	39.5%	44.5%
Total variance	1.263	2.091	100%	100%

Note: [a] Negative variance component estimates in the MINQUE method are typically due to sampling error (with the true value probably being close to zero).

The variance components in the table offer important insights into the differences between the traditional and random sampling conditions.

Respondent Variance (Var(Resp)): This variance component captures variability in the scores that is due to true variation (i.e., individual differences in PEB), but also individual variation due to memory effects and differences in scale use (including differences in acquiescence response style, which in the current set-up could not be controlled for). The proportion of respondent variance is higher for the traditional condition (14.9%) compared to the random condition (8.6%). This suggests that individual differences among respondents are more pronounced when using scales with fixed items.

Scale Variance (Var(Scale)): The scale variance component accounts for 0% of the total variance for the random condition, compared to 0.9% for the traditional condition. This zero variance in the random condition indicates that scores are not influenced by the specific scale used, reflecting the benefit of randomization in producing measurements independent of the scale. In contrast, the traditional condition's scale variance shows that different scales capture the construct in slightly different ways, introducing an additional source of undesired variability that is absent in the random condition.

Item Variance (Var(Item)): The estimate for the item variance component is substantially higher in the random condition (32.8%) compared to the traditional condition (3.0%). This higher item variance in the random condition suggests that the randomized item pool covers a broader range of the construct, thereby reducing content sampling error. In the traditional condition, the fixed set of items likely results in a narrower focus, capturing less of the construct's overall variability.

*Respondent * Time Interaction Variance (Var(Resp * Time)):* The respondent-time interaction component is low and close for both conditions (negative so presumably 0% in the traditional condition and 0.3% in the random sampling condition). This indicates that scores are relatively stable over time in both conditions, with only minor fluctuations that could be attributed to temporal factors. The slight negative value in the traditional condition is likely a result of sampling error and estimation issues rather than a meaningful effect and the true value is probably (near-)zero.

*Respondent * Scale Interaction Variance (Var(Resp * Scale)):* The respondent-scale interaction variance is higher in the traditional condition (2.9%) compared to the random condition (0.1%), indicating more scale-specific effects on individual responses in the traditional condition.

*Respondent * Domain Interaction Variance (Var(Resp * Domain)):* The respondent-domain interaction variance is significantly higher in the traditional condition (40.3%) compared to the random condition (13.7%), suggesting responses to be less domain-dependent in the random condition. As there are only two items per domain in the traditional condition (whereas there are ten per domain in the random sampling condition), part of the item-level variance might have been misattributed to the domain-level in the traditional condition, which may also help explain the relatively low estimate for the item-level variance component in the traditional condition.

Residual Variance: The residual variance is higher in the random condition (44.5%) compared to the traditional condition (39.5%). While a higher residual variance might initially suggest more unexplained

variability (which could be seen as a drawback), in this context, it likely also reflects the greater diversity of item content in the random condition.

Discussion

In this study, our objective was to provide an initial exploration of a novel approach for constructing self-report scales—the use of validated item pools and random sampling of items for each respondent. To achieve this, we compared the traditional method of administering self-report scales with the proposed randomization approach. As an illustrative example, we used the construct of Pro-Environmental Behavior in this study.

To achieve this objective, participants were randomly assigned to either the traditional or random sampling condition. Two scales were administered twice for the traditional condition, separated by a filler task. The included items for the scales in the traditional condition were also selected randomly from a larger pool (one per domain), but the same items were then used for each respondent. For the randomized condition, four different scales were created by randomly selecting one item per domain from a larger item pool for each respondent. The four scales were also separated by a filler task to make it comparable to the traditional approach (with two scales being administered before and two after the filler task).

By correlating the sum scores of the different scales, we could investigate the convergent validity. Correlations in the randomized condition are generally lower than those in the traditional condition. This discrepancy largely stemmed from differences in scale composition; unlike the traditional condition where scales consist of the same items measured at different times, scales in the randomized condition feature different items, leading to lower raw correlations. Moreover, the results of a variance decomposition provide more in-depth insights. Some of the variance components we distinguished will generally positively contribute to the scale correlations (reported in Table 1), some will not affect them, and others

will negatively affect them. To clarify, we will now, in turn, discuss three sets of variance components: those that involve the respondent facet, those that involve the item facet, and finally the scale facet.

Let us consider variance components involving the respondent facet first. The scale correlations (see Table 1) were computed using respondents as the unit of analysis. Therefore, the respondent-level variance term (Var(Resp)) positively contributes to these correlations. This component was larger in the traditional condition than in the random condition. It is not completely clear why this is the case and there may be several contributing factors. For one, consistency over time, partly due to memory effects, may contribute to this variance component. Another part of the reason is probably that the random condition contains more respondent-by-item interaction information (discussed later). The interaction terms between respondent and time, and between respondent and scale, negatively affect the scale correlations (as they correspond to variation within respondents over time or scales), but these components were relatively small. That being said, we note that the respondent-by-scale interaction accounts for slightly more variation in the traditional condition. As each scale contains one item per domain (note that domain was treated as a fixed factor), variation due to the domain-by-respondent interaction will generally not affect the scale correlations. The domain-by-respondent interaction variance was lower for the random condition, suggesting that responses are more uniform across domains, likely due to the broader content coverage (and thus lower content sampling error) within a domain provided by random item selection.

Next, let us consider the variance components involving the item facet. As the items are constant across respondents in the traditional method, their variance does not affect the scale correlations. This is not the case for the random condition, where different respondents randomly get different items (e.g., respondent 1 might answer “I avoid using electricity during the peak period (8.00 a.m. to 10.00 p.m.)”, whereas respondent 2 might answer “I turn off lights when not in use”). Thus, in the random condition, item variation is confounded with respondent variation when analyzing the data using only respondents as the unit of analysis (as is the case for the correlations in Table 1). In addition, variation due to the

interaction between respondent and item cannot be estimated independently from random variance; both random variance and the respondent-by-item interaction variance are therefore included in what is called the residual variance in Table 2. As a rule, residual variance negatively affects correlations like the ones in Table 1 (Baumgartner et al., 2021). Overall, between-item variation helps explain the lower scale correlations in the random condition compared to the traditional condition (in Table 1). However, the observation that the random condition shows more item-related variation also supports the idea that it leads to better content coverage.

Lastly, we consider the scale variance component. The scale variance is zero in the random condition, highlighting the advantage of randomization in producing measurements that are independent of the specific scale used (as essentially, there are no different scales, just different item samples). In contrast, the traditional condition exhibits some scale variance, indicating that different scales capture the construct in slightly different ways, possibly due to content sampling error. This introduces an additional source of undesired variability that is absent in the random condition. Such scale variance might reduce the correlation between the two traditional scales, as each scale introduces a unique source of variance, leading to greater divergence in the scores.

Overall, the results indicate that the randomization of items may be effective in minimizing content sampling error, as suggested by the higher item variance and lower scale-specific and domain-specific interaction effects. While the random condition shows higher residual variance, this variance likely includes both meaningful variation related to the construct and random error. The traditional condition, with its higher respondent and scale variances, seems more prone to limited content coverage, leading to limited scope and comparability of the scales. The lower correlations between scales in the random condition are primarily due to higher item and residual variance components compared to the traditional condition, combined with a lower respondent variance. This highlights a trade-off between achieving consistency across scales and reducing content sampling error.

Going beyond our empirical exploration, the random item sampling method may offer several advantages compared to the use of validated self-report scales with fixed items. First, if all researchers systematically build their scales within the same predefined domain framework and by sampling from the same validated item pools, then measures may be more comparable between different studies and may have higher chances of covering the full construct domain, thus increasing content validity. This could be accomplished by making these item pools available on open science repositories such as Open Science Framework or self-developed item repositories.

Second, by randomly selecting items per domain for each respondent, systematic content sampling errors may be accounted for as it may balance itself out on an aggregate level. This, of course, depends on the quality of the item set from which the sampling is done. Moreover, it lifts content sampling error and representative sampling from the universe of items from a theoretical idea to a more concrete part of measurement.

Third, the use of a validated item pool may enhance validity as the researcher might also discard items that are not valid for the context and sample under study, leading to the flexible inclusion of items. While we utilized random sampling of items for each respondent in this study, researchers also have the option to sample a predetermined set of items per domain (purposive sampling). Although this method might not address content sampling error directly as all respondents need to respond to the same set of items, it still necessitates the researcher's active engagement in item selection. This involvement can contribute to content validity, as the chosen items may be tailored to fit the specific sample and context of the study.

Two sidenotes should be addressed, including (1) determining whether the construct is reflective, formative, or a combination of both and (2) establishing the method for defining the domain structure.

Related to the first side note, in the case of reflective constructs, the causality flows from the construct to the items, indicating that items merely reflect the underlying construct. This type of construct is independent of its indicators. Conversely, formative models determine that there is a logical relation between the indicators and the construct, without necessarily making any claims on causality. Formative constructs are employed for their functional utility. They serve as tools for the prediction of other variables or facilitate the exploration of interventions. For instance, one might seek to understand how the composite frequency score of Pro-Environmental Behavior (PEB) changes with the implementation of specific interventions. In the context of this study, the construct of Pro-Environmental Behavior (PEB) was defined as a combination of a formative and reflective construct. This is because behavioral items may not distinctly be caused by only one underlying motivator. For instance, the response to an item like 'I cycle to work' could for respondents be differentially caused by monetary sensitivity (e.g., receiving compensation per kilometer driven or being sensitive to high fuel prices), general fitness, or time sensitivity (e.g., being faster at work with a bike than with another mode of transportation). However, within a behavioral domain, items might correlate as they are considered more homogeneous. For example, someone who cycles to work might also cycle a lot in general, causing less car use. An item representing car use will therefore correlate with an item representing cycling behavior, with both behaviors being part of the transportation domain. The random sampling method is only applicable to reflective measures or the reflective component (such as in our example) of a formative-reflective measurement model.

In connection with the second side note, multi-dimensional constructs can be conceptualized using distinct domain models. For example, Fried (2017) discovered that seven widely-used depression scales exhibited low content overlap. This heterogeneity stemmed from authors' varying conceptualizations of depression; some viewed it as a brain disorder, others as a clinical form of grief, or as a set of self-defeating attitudes. Similarly, in the context of Pro-Environmental Behavior (PEB), domains may be influenced by different perspectives. For instance, domains could be operationalized by factors

such as the difficulty of engaging in certain behaviors, their level of environmental impact, or their similarity (e.g., transportation-related behaviors). Consider two researchers attempting to measure PEB: one might select items based on difficulty, encompassing behaviors of varying complexity (e.g., adherence to a vegetarian diet vs. turning off lights when leaving a room). Meanwhile, another researcher might focus on behavioral similarity, including items related to transportation, recycling, and energy use (as utilized in this study). Consequently, self-report scales intended to measure PEB may yield different outcomes, reflecting the emphasis on different features of the behaviors included. Therefore, it is crucial to clearly define the domain structures when utilizing self-report scales. Note that a construct can also be conceptualized as unidimensional, stemming from a single domain. However, the presence of different domain structures may not necessarily pose an issue, as they can be tailored to address specific research questions. Hence, selecting domain structures that align with the research objectives is imperative and should always be clearly discussed by researchers.

Limitations

In this study, we only included one item for each domain. This might have had an influence on the reliability within the domains. As such, future studies may include more items per domain to account for more random measurement error within the domains, which might in turn increase the correlations between the outcomes of two randomized scales (hence increasing the reliability). The optimal number of indicators per domain is a research question on its own. However, as a general rule of thumb (and when using standard parameterization approaches), a latent factor needs three indicators for a standalone factor to be identified, and four indicators for it to be overidentified, which is generally desirable (Baumgartner, & Weijters 2019).

Second, we used Pro-Environmental behavior as an example here, it is thus important to also test this procedure on other constructs. Third, the pertained advantages of controlling for systematic content

sampling error, need more thorough psychometric investigation in future studies. Yet, the theoretical nature of the universe of potential items and the challenge of representative sampling from it impose limitations on conducting empirical studies.

Conclusion

The use of validated item pools with random item sampling emerges as a promising avenue to consider for studies utilizing aggregate data. Departing from traditional validated scales with fixed items presents a potential remedy to the persistent challenge of content sampling error and validity problems and may offer several advantages. First, random item sampling minimizes content sampling error by making it random rather than systematic, averaging it out across studies. Second, it enhances consistency and comparability across studies by reducing structural and content differences typical for traditional scales with fixed items. Third, it prompts researchers to engage more critically with item selection, ensuring items are contextually valid and reducing over-reliance on standardized scales. Finally, validated item pools are easily updated, allowing for the dynamic application of self-report scales without needing to revalidate entire scales. With the ideas presented in this paper, we hope to inspire further investigation into this methodology.

References

Antonakis, J., & Bastardoz, N., Jacquart, P. & Shamir, B. (2016). Charisma: An Ill-Defined and Ill-Measured Gift. *Annual Review of Organizational Psychology and Organizational Behavior, 3*, 293-319.

Baumgartner, H., & Weijters, B. (2019). Measurement in Marketing. *Foundations and Trends® in Marketing, 12*(4), 278-400.

Baumgartner, H., Weijters, B., & Pieters, R. (2021). The biasing effect of common method variance: Some clarifications. *Journal of the Academy of Marketing Science*, *49*, 221-235.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061–1071.

Campbell, S., Greenwood, M., Prior, S., Shearer, T., Walkem, K., Young, S., Bywaters, D., Walker, K (2020). Purposive sampling: complex or simple? Research case examples. *Journal of Research in Nursing, 25*(8), 652-661.

Churchill, G. A. J. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research*, 64-73.

Cook, T. D., Campbell, D. T., & Day, A. (1979). *Quasi-experimentation: Design & analysis issues for field settings* (Vol. 351): Houghton Mifflin Boston.

Cote, J. A., and Buckley, R. (1987). Estimating trait, method, and error variance: generalizing across 70 construct validation studies. *Journal of Marketing Research, 24*, 315–318.

Cronbach, L. J. (1949). Essentials of Psychological Testing. Harper.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles. New York: Wiley.

Deltomme, B., Gorissen, K. & Weijters, B. (2023). Measuring Pro-Environmental Behavior: Convergent Validity, Internal Consistency, and Respondent Experience of Existing Instruments. *Sustainability, 15*(19), 14484.

Edwards, J. R. (2010). The fallacy of formative measurement. *Organizational Research Methods*, *14*(2), 370-388.

Flake J. K., Fried, E. I. (2020) Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456-465.

Fried E. I. (2017). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *Journal of Affective Disorders*, *208*, 191–197.

Fried, E. I. & Flake, J. K. (2018). Measurement Matters. The Observer, published by the Association for Psychological Science.

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). Survey methodology. John Wiley & Sons.

Howell, R. D., Breivik, E., & Wilcox, J. B. (2007). Reconsidering formative measurement. *Psychological Methods, 12*(2), 205.

Jarvis, C. B., MacKenzie, S. B., & Podsakoff, P. M. (2003). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of consumer research, 30*(2), 199-218.

Koppold, A., Ruge, J., Hecker, T., & Lonsdorf, T. B. (2023). The many faces of early life adversity - content overlap in validated assessment instruments as well as in fear and reward learning research. PsyArXiv.

Lilienfeld, S. O., Strother A. N. (2020). Psychological Measurement and the Replication Crisis: Four Sacred Cows. *Canadian Psychology, 61*(4), 281-288.

McDonald, R. P. (2003). Behavior domains in theory and in practice. *Alberta Journal of Educational Research*, 49(3).

Markus, K. A., & Borsboom, D. (2013). Frontiers of test validity theory: Measurement, causation, and meaning. Routledge/Taylor & Francis Group.

Meier, S. T. (1994). *The Chronic Crisis in Psychological Measurement and Assessment.* New York, NY: Academic Press.

Meier, S. T. (2008). *Measuring Change in Counseling and Psychotherapy.* New York, NY: Guilford Press.

Meier, S. T. (2023). Editorial: Persistence of measurement problems in psychological research. *Frontiers in Psychology, 14.*

Nunnally, J. C. (1967). Psychometric Theory. New York: McGraw-Hill.

Rhemtulla, M., van Bork, R., & Borsboom, D. (2020). Worse than measurement error: Consequences of inappropriate latent variable measurement models. *Psychological Methods, 25*(1), 30.

Shoemaker, D. M. (1973). Principles and procedures of multiple matrix sampling. Ballinger.

Thurstone, L. L. *(*1927*).* A law of comparative judgment. *Psychological Review, 34*(4), 273–286.

Warnell, K. R., & Redcay, E. (2019). Minimal coherence among varied theory of mind measures in childhood and adulthood. *Cognition*, *191*, 103997.

Weidman, A., Steckler, C. & Tracy, J. (2016). The Jingle and Jangle of Emotion Assessment: Imprecise Measurement, Casual Scale Usage, and Conceptual Fuzziness in Emotion Research. *Emotion*, *17*(2), 267-295.