Academic Research Values Scale:

Item selection and content validity

Andrea Kis[1], Tatiana Marci[2], Gianmarco Altoè[2], Flavio Azevedo[3], Elena M. Tur[1], Wybo Houkes[1], Daniël Lakens[1]

[1] Department of Industrial Engineering & Innovation Sciences, Eindhoven University of Technology

[2] Department of Developmental Psychology and Socialisation, University of Padova

[3] Department of Interdisciplinary Social Science, Utrecht University

Author Note

Andrea Kis - https://orcid.org/0000-0002-4345-3814

Tatiana Marci – https://orcid.org/0000-0002-2813-0312

Gianmarco Altoè – https://orcid.org/ 0000-0003-1154-9528

Flavio Azevedo - https://orcid.org/ 0000-0001-9000-8513

Elena M. Tur - https://orcid.org/0000-0001-9634-0090

Wybo Houkes - https://orcid.org/0000-0003-3148-4805

Daniël Lakens - https://orcid.org/0000-0002-0247-239X

We have no conflict of interest to disclose.

Correspondence concerning this article should be addressed to Andrea Kis, Eindhoven University of Technology, 5612 AZ Eindhoven, The Netherlands. Email: a.kis@tue.nl

Abstract

In this paper we test the content validity of 246 academic research value items in two interrelated methodological steps. First, we analyze the in-depth evaluation of two experts to assess formal aspects of our items and their relatedness to the construct. Next, we evaluate the relevance of each item based on feedback gathered from 20 experts. Based on these assessments, we review and refine the original item pool and propose a total of 97 value items spread through 10 dimensions. We relate our items to existing work on researchers' values and provide recommendations for future measurement development. A validated measure of academic research values can assess the effectiveness of responsible research conduct courses and highlight personal differences among researchers from diverse backgrounds, as well as aid our understanding of values unique to academic careers.

Keywords: academic research values, content validity, scale development, personal values, work values

Academic Research Values Scale: Item selection and content validity

Introduction

Values are desirable goals that serve as guiding principles and transcend specific situations (Schwartz, 1992; Schwartz et al., 2012). They are psychological constructs considered central to understanding human behavior (English et al., 2018) and increasingly cited as necessary instruments for social and ethical judgements in science (Douglas, 2023) as well as precursors for equitable and sustainable operationalization of (open) scientific standards (UNESCO, 2021). Values play a prominent role in terms of their importance for sustainable scientific development and the coordination of scientific research, as well as in improving our understanding of some of the mechanisms that lead to exemplary and questionable science. Yet, there is a lack of valid and reliable tools to measure values in academic research environments (Kis et al., 2023).

This represents an important gap in light of the credibility crisis facing the social sciences: recognizing and assessing the role of values might provide important insights for understanding and promoting integrity, transparency, and accountability in scientific research (English et al., 2018). This lack of measurement tools may reflect – at least in part – the shortage of well-accepted and formalized value theories. In turn, it might also signal the challenges of ensuring adequate content coverage of this multifaceted construct's indicators. Therefore, the aim of this study – as part of a broader construct validation process – is to examine the content validity of a set of academic value indicators (items). These items are intended as a basis of a theory-driven and psychometrically robust instrument for evaluating values in academic research settings.

In a previous study, we conducted a review of the personal, work, and scientific values literature (Kis et al., 2023). Based on this review we defined values in academic research settings as "principles which serve as a basis of evaluating outcomes of scientific

work-related actions, guide the selection of scientific work goals, and represent the relative importance assigned to various academic job aspects related to research activities" (Kis et al., 2023, p. 23). We defined academic research values by integrating the values literature with our own findings from interviews and surveys with researchers. Then, we operationalized them by developing an initial pool of 246 academic research value items, spread across 36 sub-themes in 11 dimensions: Ambition, Authority, Autonomy, Benevolence, Conformity, Enjoyment, Organizational support, Tradition, Universalism, Variety, and Working environment. For an overview of definitions and example items, see Table 1. An extended explanation of the whole procedure is provided by Kis et al. (2023).

Documenting the development steps rigorously is required to ensure the creation of a solid basis for a scale measuring academic research values, especially given the shortage of existing measurement tools and the benefits of studying values in science. Values are widely accepted to influence behavior (Sagiv et al., 2017; Sagiv & Roccas, 2021), making them useful tools to improve our knowledge of how values affect research practices. For example, we might be able to find differences between the researchers who commit to responsible conduct of research and those who engage in questionable practices. A validated academic research value measure can help evaluate the effectiveness of responsible research conduct courses and explore personal differences among researchers from various nationalities, disciplines, and career stages (English et al., 2018). In addition, understanding values unique to research environments can increase the appeal of academic careers to a diverse group of scholars.

Table 1

Academic research value dimensions, definitions, and examples of items

Dimensions	Definitions	Example items
Ambition	Career success through demonstrating competence according to academic standards, feelings of achievement and being a competent researcher	To win grants, scholarships, and scientific awards To be highly cited
Authority	Scientific status, wealth, and prestige, control or dominance over other researchers and research resources, the importance of having a good public image as a researcher	To make decisions about who does what in a research project To have direct influence over funding decisions To lead a prestigious research group
Autonomy	Freedom of thought and action: determination of work tasks, creating, and exploring own research topics	To be able to set my own research agenda To determine how I spend my workday
Benevolence	Being committed to the welfare of other researchers and emphasizing the importance of dependability and relationships within the research community	To help the people in my research community To have good interactions with fellow researchers
Conformity	Conformity to scientific norms and codes of conduct, restraint of actions that might upset or harm others, abiding by social norms within the work environment	To work with researchers who respect scientific norms To return favors to collaborators and colleagues
Enjoyment	Seeking to take pleasure and gratification within the realm of scientific work, enjoy doing research	To go on nice conference trips To enjoy my work
Organizational support	Fairness, support, and clarity within the research organization	To know that the research institution handles processes fairly To feel supported by the university I work at
Tradition	Modesty about achievements and role as a researcher, respect, and acceptance of scientific traditions	To do scientific work which would be traditionally approved of To be modest about my scientific achievements
Universalism	Assigning importance to research that has a positive social impact, sense of need to contribute to sustainability and prevent unethical or immoral research behaviors, and being tolerant to different approaches	To better the world with my research To make sure that the outcomes of my research do not have harmful consequences for nature To protect scientific integrity
Variety	Being drawn to innovation, variety, novelty, and challenge in research, emphasizing the importance of personal growth and learning	To do varied work To encounter exciting new ideas To uncover hidden truths
Working environment	Personal safety and comfort within the working and broader scientific environment, a sense of job security	To work in an environment free from abusive relationships To have a job that provides steady employment

In the current study, we focus on testing the content validity of the generated items in two interrelated steps. First, two experts provide an in-depth evaluation of the items' formal aspects (grammatical and syntax validity) and the extent to which each relates to the construct for which it was designed (face validity). Based on the results of this first step, we assess their content validity by evaluating the relevance of each item on the basis of the dimension for which it was designed by a second, larger group of experts (N = 20). In addition to assessing each item's readability, clarity, and relevance, this validation phase allows us to review and refine the initial pool of items before moving forward to further exploratory analyses. This article reports on the results of this process with the aim of rigorously and transparently disclosing content validity assessments and facilitating a discussion about proposed value items prior to proceeding with large sample scale development steps.

Background

We continue the psychometric measurement development started by Kis and colleagues (2023). Publishing these studies as a standalone paper – despite being substeps of a psychometric development process – has two main benefits. First, by publishing these studies independently, we highlight how much effort needs to be invested in each substep of the psychometric research process. Scale development relies on complex and systematic procedures that require a rigorous approach to all stages of the process (Boateng et al., 2018; Morgado et al., 2017). The same rigor in documenting associated methodological choices and steps is essential for maintaining scientific integrity, as well as ensuring transparency and thus enabling other researchers to scrutinize, replicate, or build upon our results. Second, well-documented theoretical analysis of content validity assessments within psychometric work are often missing. Some studies fail to report whether and how they analyzed the opinions of experts and their target population, leaving readers uncertain as to whether only the report or the step itself is missing (Morgado et al., 2017).

Construct validation and content validity

The concept of validity has undergone several conceptualizations and definitions over time (American Educational Research Association et al., 2014; Anastasi, 1950; Haynes et al., 1995; Messick, 1989). The current view is largely based on Messick's (1995) framework (Brown, 2010). Messick proposed a comprehensive concept of validity where various aspects - content validity, criterion-related validity, construct validity, and consequential validity - are unified. They defined it as a general judgment about how well evidence and theories support the interpretation and actions based on test scores or other assessment methods (Messick, 1995, p. 741).

Content validity – the adequacy with which a measure evaluates the area of interest – has been pointed out as a prerequisite for testing other aspects of construct validity (Koller et al., 2017). Although existing guidelines highlight the importance of content validity within the construct validation process (American Educational Research Association et al., 2014; Flake et al., 2017), it is often overlooked in practice, and there is a widespread tendency to assume that content validity is limited by the theoretical definition of the construct (Koller et al., 2017), leading to potential subsequent problems. For example, the items may encompass only parts of the construct or may not reflect the intended construct. If content validity is not accomplished, there is no point in testing other aspects of validity. The process of content validity involves several components, such as the formal aspect of the items (e.g. clarity and linguistic aspects of the items), the validity and representativeness of the construct definition, and the appropriateness of the response format (Koller et al., 2017). It includes both qualitative and quantitative approaches and is mainly assessed through the judgment of experts and members of the target population (Boateng et al., 2018).

The specific aims of the pre-validation phase presented in this paper are: a) to ensure the formal adequacy and face validity of each item (Step 1), assessing whether the items are

clear, understandable, grammatically correct and appear to measure what they are intended to measure; b) to identify the most relevant related construct items (Step 2), integrating a quantitative assessment based on the Content Validity Index (Lawshe, 1975; Polit & Beck, 2006) with expert insight.

Methods

Content validity assessment: Expert and target population involvement

In line with recommendations and best practices (American Educational Research Association et al., 2014; Boateng et al., 2018), we based our content validity assessments of the 246 value items derived from our previous study (Kis et al., 2023) using both quantitative and qualitative methods across two interrelated steps involving two panels of experts (Step 1: N = 2; Step 2: N = 20). We selected experts who are also members of our target population of researchers. A step-by-step overview of all methods used during the process and main results is presented in Table 2.

Step 1 - Formal aspects and face validity: mixed methods evaluation by experts

To assess formal appropriateness (grammatical, syntax, and comprehensibility) as well as face validity of each initial value item, we invited two experts from our direct academic environment in August 2023. In line with (Haynes et al., 1995), in our study we will refer to face validity as "a component of content validity" namely the "degree that respondents or users judge that the items of an assessment instrument are appropriate to the targeted construct and assessment objectives (Allen & Yen, 1979; Anastasi, 1988; Nevo, 1985)" (p. 243). This preliminary step follows a previous evaluation within the research group (as explained in Kis et al., 2023) and is followed by the evaluation of a larger number of experts (Step 2 of the current manuscript). Given the goal of our inquiry, the information richness of the input of our experts, and the analytical steps, we agreed that the sample size does not have to be large.

Eligibility was based on language skills (i.e., native English speakers), disciplinary expertise with meta-scientific inquiries (e.g., philosophers of science or linguistic experts), as well as independence from this line of research and availability. Eligible experts were invited by members of our research team and asked to self-administer an online survey containing a description of their task, all 246 initial value items and definitions of connected sub-themes and dimensions (see the project's OSF repository). No compensation was offered to these experts.

Table 2

Step-by-step overview of the process to develop the Academic Research Values (ARV) scale

Steps	Aim	Methods	Results
Item selection and content validity: Assessing if the items adequately measure the domain of interest			
1. Formal aspects and face validity evaluation by experts	To evaluate each item in terms of formal appropriateness (grammatical, syntax, and comprehensibility) as well as face validity	Experts (N=2): Native English speakers, philosophers of science Evaluation of each item: 1) link to dimension (Yes / No) 2) clarity of item (1 (low) / 2 / 3 (high)) 3) suggestions (text) 4) overlap with other items (text) + other comments (text)	Quantitative and qualitative assessment of expert comments lead to reducing the list by 41 items (N = 205), see Figure 1.
2. Content validity index by experts	To evaluate each item in terms of relevance	Experts (N=20): Researchers from all ranks and disciplines involved in the work of academic organizations Evaluation of each item: Relevance rating of each item (item not relevant / somewhat relevant / quite relevant / highly relevant to the category it is displayed in, to researchers in general)	Quantitative and qualitative assessment of responses and theoretical considerations underlying value dimensions and sub-themes led to reducing the list by 105 items (N = 98), see Figure 2.
Survey administration and sample size: gathering enough data from the right people - *Future research*			

For each item, we asked the experts to provide four types of feedback. In terms of formal appropriateness, experts were asked to evaluate the extent to which the item is comprehensive and clear on a 3-point Likert scale (from 1 = low, to 3 = high) and to provide their suggestions for improving items (e.g., for reformulation, noting redundancies, noting issues / difficulties with dimension definitions). Experts also reported the presence of redundant items (i.e., items representing the same concept). To assess face validity, experts judged if the displayed items are linked (yes or no), to the dimensions they are presented in. In addition, experts opted to also share general feedback about the questionnaire.

Step 2 - Content validity index: quantitative evaluation by experts

As part of quantitative content validity evaluation, multi-rater kappa index (Kappa) and Content Validity Index (CVI) were assessed based on the responses of a panel of experts working in the Netherlands. The project was registered in the study proposal and ethical form approved by the TU/e's Ethical Review Board (ID: 1914, see OSF) prior to data collection between November and December 2023. For this stage of the process, we conceptualized experts as researchers of all ranks who have a keen interest in and expertise with discussions about the values and future of science and academia. Accordingly, eligibility and invitations were based on membership or active participation in organizations aimed at bettering science and academia as a whole (e.g., the Young Academy of the Netherlands and Promovendi Netwerk Nederland, the national interest group for and by PhD candidates).

To represent a wide range of academic perspectives and to align with our prior value studies, we aimed to include researchers across all disciplines regardless of age or nationality while including an equal number of career stages, aiming to keep gender equality. There is still no agreement in the literature on the number of experts required for content validation (see Roebianto et al., 2023 for an overview) and the proposed number of experts varies from a minimum of three (Lynn, 1986) even up to twenty (Gable & Wolf, 1993; Rubio et al.,

2003). Some authors recommended caution in rigidly adhering to specific numerical thresholds, advocating instead for flexibility that takes into account the complexity of the construct being measured and the expertise required (Almanasreh et al., 2019; Grant & Davis, 1997). At least five people are recommended to have sufficient control over chance agreement (Zamanzadeh et al., 2015). To sum up, while there is no consensus on the specific number of experts required for content validation, the literature points out that a balanced approach that integrates quantitative guidelines with qualitative considerations is crucial to ensure the validity of assessment instruments.

For our self-report survey, we require the experts to encompass the diverse perspectives and expertise within the academic community. Specifically, the variables we have considered are academic position (PhD or postdoc vs. assistant, associate, or full professors) and scientific discipline (STEM vs. non-STEM). We therefore decided to recruit a total of 20 experts: 5 PhDs or postdocs in STEM disciplines, 5 PhDs or postdocs in non-STEM disciplines, 5 professors in STEM disciplines and 5 professors in non-STEM disciplines.

Eligible experts working in the Netherlands were invited via their organizational email addresses, or social media channels and mailing lists operated by the organizations. They were invited to participate in a self-administered online survey containing a description of their task, all value items remaining after Step 1, and definitions of connected dimensions (see OSF). A compensation of 50€ (PhD candidates and postdocs) or 100€ (professors) was offered to experts for their time in their chosen format (donation / gift card / bank transfer).

Consistent with extant practice (Polit & Beck, 2006), we asked the experts to evaluate the relevance of each value item to the category it was displayed in, using a 4-point Likert scale (1 - the item is not relevant to the category it is displayed in; 2 - the item is somewhat relevant to the category it is displayed in; 3 - the item is quite relevant to the category it is

displayed in; 4 - the item is highly relevant to the category it is displayed in). They were instructed to be objective and constructive, to rate each item using the entire scale, be sensitive to differences between the four options, and to evaluate the construct as one about researchers in general rather than themselves – values that might be of importance to at least some researchers even if they might not be as vital to others or the experts themselves.

Based on the literature (see Lynn, 1986; Shrotryia & Dhanda, 2019), responses to each item were dichotomized as 0 (not relevant) for "somewhat relevant" or "not relevant" responses and 1 (relevant) for "relevant" or "quite relevant" responses. To assess the validity of each item, the content validity index was calculated as the proportion of 1 (relevant) to the total number of ratings (i.e., 20 = number of experts). Quantitative procedures for evaluating content validity, including the use of cutoff values, have been proposed to encourage a more systematic approach to the content validation process. These procedures offer valuable guidelines for assessing the degree of agreement among experts regarding the relevance and representativeness of assessment content. However, these cutoff values are subject to adjustments across studies (Hardesty & Bearden, 2004) and should be interpreted based on the nature of the construct, target population, and intended assessment use (Messick, 1995). It is also recommended to use quantitative procedure in association with qualitative insights from experts (Delgado-Rico et al., 2012; Spoto et al., 2023) that may provide judgments on contextual factors (see Haynes et al., 1995) that may not be captured by quantitative measures alone.

Although the CVI is extensively used to estimate content validity, this index does not consider the possibility of inflated values because of the chance agreement. To address this shortcoming, we also considered the multi-rater kappa index (Kappa) because, unlike the CVI, it suitably adjusts for this potential biased chance agreement. Based on the extant literature, values below .40; between .41 and .60; between .61 and .80; and above .80 were

considered to indicate fair, moderate, good, and excellent agreement respectively (Polit & Beck, 2006; Viera & Garrett, 2005).

Accordingly, during initial evaluation, items with a Kappa index lower than .60 were categorized as "not appropriate". All other items (i.e., items with a Kappa index greater than or equal to .60) were rated based on their CVI, labeling items either "appropriate" (CV I ≥ .78), "to be revised" (.70 ≤ CV I ≤ .77), or "not appropriate" (CV I ≤ .70). These labels were based on the existing literature recommending the cut-offs to be set according to the number of experts – with cut-off limits decreasing as the number of experts increases. In the literature, when there are more than 9 experts, researchers tend to use .78 as the cut-off value. However, it is plausible that in our case of 20 experts, the cut-off should be lowered further and .70 might be considered a reasonable cut-off value.

Results

Step 1 - Formal aspects and face validity: mixed methods evaluation by experts

In Step 1, items were pre-validated by two philosophers of science. Results were assessed using a mixed-methods approach as described in Table 3. From the complete pool of 246 items, a number of 122 items reported problems connected to clarity (the extent to which the item is comprehendible), link to dimensions (whether the item is linked to the corresponding dimension), and other aspects. These other aspects were categorized based on experts' suggestions as issues of ambiguity, broadness or vagueness, relevance, phrasing, miscategorization, connection between question formulation and item, and peculiarity. Items with problems were then grouped based on ease of correction (easy, moderate, hard) to facilitate decisions about inclusion. After rephrasing items with phrasing difficulties and excluding the most problematic items, the 240 item list was further reduced to 212 items based on experts' suggestions about overlaps and redundancies. In cases of redundancy, the more formally appropriate item would be chosen. After discussing cases labeled hard to

correct and a final round of overlap-related exclusions, this step resulted in a list of 205 value items, as depicted in Figure 2. All data and analysis files as well as a more extensive description of methods and results are available in the project's OSF repository.

Figure 1

Decision tree for Step 1

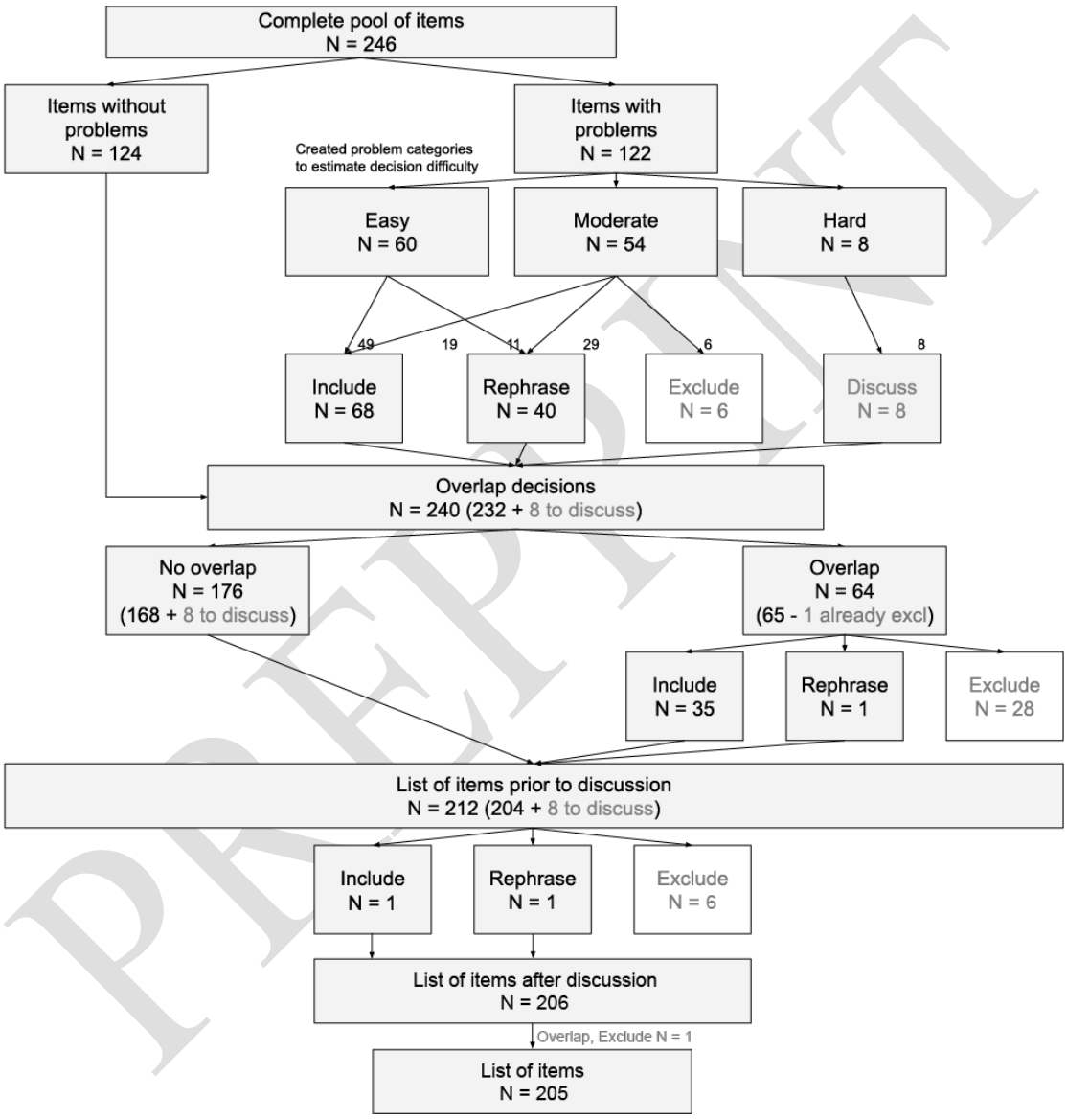


Table 3

Step 1 step-by-step analysis of results

Method	Response type analyzed	Description	Main results
Quantitative	Link to dimension	Description of the number and rate of items with reported problems, also as a factor of dimensions.	88% of items were considered to be linked to their dimension by both experts, 0% unlinked, with the Universalism (70%), Authority (82%), and Variety (81%) dimensions being the lowest rated.
	Clarity		Both rated 52% clear (clarity = 3), 48% of items marked less (clarity < 3) by at least one. Lowest rated (at least one expert < 3): Organizational support (83%), Authority (82%), Ambition (77%).
Quantitative	Link to dimension	Description of number/rate of items without problems, to separate items with no problems and items with any detected problems.	124 items (50%) were rated both clear and linked to the dimension they were listed in. Best dimensions (rate of "no problem" items): Autonomy (100%), Tradition (92%), Working environment (89%).
	Clarity		
Quantitative	Link to dimension	Description of number/rate of items with problems (with link to dimension or clarity).	1 problem: 37% of items 2 problems: 11% of items 3 problems: 2% of items (0% items with 4 problems)
	Clarity		
Qualitative	Suggestions	Expert suggestions on problematic items were reviewed and coded using thematic analysis. Steps: development of initial coding, identification and review of codes, and defining themes, as described by Braun and Clarke, 2006.	Problem-related suggestions were categorized into 8 types: ambiguity, broad/vague, category, phrasing, question, relevance, peculiar (+ N/A, when no suggestions were available). A total of 136 suggestions were reviewed.
Qualitative	Link to dimension		Easy: 6 subtypes, 184 items (incl items with no problems)

	Clarity	Evaluating the difficulty of correcting items - each item was categorized based on the types and number of problems detected (Easy, Moderate, or Hard). Then, each received a recommendation (include, rephrase, exclude, or discuss).	Moderate: 4 subtypes, 54 items Hard: 1 subtype, 8 items
	Suggestions		Include (192 items, 78%) Rephrase (40 items, 16%) Exclude (6 items, 2%) Discuss (8 items, 3%)
Qualitative	Suggestions	Expert suggestions on overlaps were reviewed one by one.	A total of 65 overlaps were handled. Suggestions: include (35 items, 14%), rephrase (1 item, 0%), exclude (28 items, 11%), already excluded (1 item, 0%).
	Overlap with other items		
Quantitative	All expert response types analyzed so far	Summary of decisions per dimension and sub-theme.	A total of 34 items were eliminated (-13.82% change) and 8 items were marked for discussion with the larger team.
Qualitative	Suggestions	All expert suggestions on overarching issues (e.g., construct and dimension definitions, interpretation of dimensions) and other comments were reviewed.	14 comments (E1: 11, E2: 3) were reviewed, of which 6 were deemed important to discuss.
	Other comments		
Qualitative	All response types	All responses were considered to finalize the list of items.	Of the 8 items marked for discussion, 6 were excluded, 1 was included, 1 was rephrased. All remaining questions and comments were resolved.
Qualitative	List of items	The list of items was checked again before Step 2 for inconsistencies.	1 item was excluded due to overlapping meaning with another item.

Step 2 - Content validity index: quantitative evaluation by experts

Based on the outcomes of Step 1, multi-rater kappa index and content validity index were calculated at item and scale level based on 20 experts' responses to 205 items. Because 21 experts (5 PhD candidates in STEM and 5 in non-STEM disciplines, 6 professors in STEM and 5 professors in non-STEM disciplines) finished filling out the survey due to the invitation methods, we excluded 1 professor randomly from the STEM group prior to analysis to equalize the number of participants per group. This was done to ensure that each group had the same weight in determining the multi-rater kappa index and content validity index.

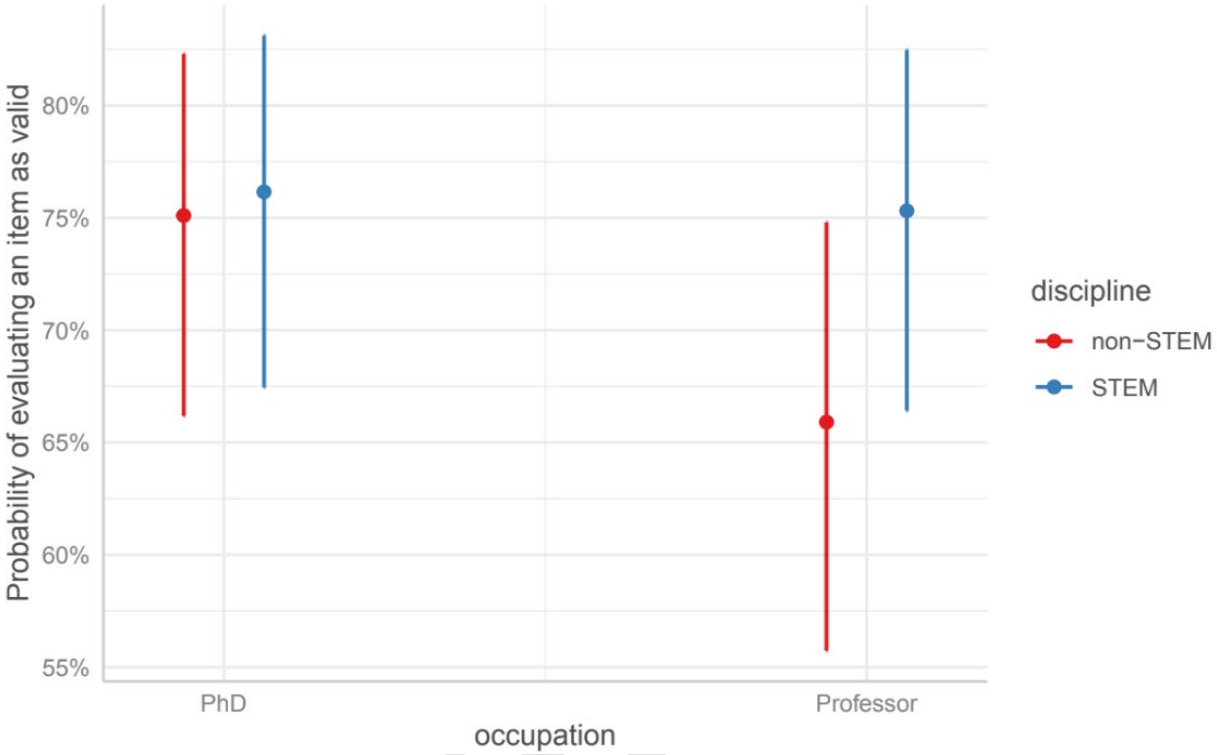
Content validity index calculation

No missing data was present. One item was also excluded based on redundancy (i.e., almost completely overlapping phrasing) after analysis, keeping the item with the higher validity. Namely, in the Benevolence dimension, "To feel like I am a part of the research community" was included and "To feel that I am a part of the research community" was excluded.

To assess whether the probability of evaluating an item as valid was associated with academic status (PhD/postdoc vs. professor), discipline (non-STEM vs STEM), or a factor of an interaction between academic status and discipline, a logistic mixed-effects regression was conducted. Random effects included random intercepts for Subjects (N = 20) and Items (N = 204). From the analysis of model deviance, no effect was found to be statistically significant at the 5% level: academic status ($\chi^2(1) = 1.430$; $p = .232$), discipline ($\chi^2(1) = 1.560$; $p = .212$), interaction ($\chi^2(1) = .937$; $p = .333$). Although there was no statistically significant effect of academic status and discipline, the small sample size does not allow us to detect small differences, and descriptively, there might be a slightly lower score from non-STEM professors (see Figure 2 for descriptive results).

Figure 2

Estimated effects based on means and estimated 95% confidence intervals



As can be seen in Table 4, the results obtained after categorizing items based on CVI values are not uniform across the different scales. While items contained within the Organizational support and Working environment dimensions were (almost) all substantially related to their theoretically proposed factors, the Tradition dimension almost completely vanished, with only one item reporting a satisfactory CVI rating. For a reflection on how this outcome relates to the broader literature and our prior findings, see the discussion.

Table 4

Summary of outcomes based on CVI per dimension

Dimension	$N_{initial\ items}$	Yes	Revise	No	$N_{final\ items}$	Change	No %
Ambition	22	5	7	10	12	-10	45%
Authority	29	3	7	19	10	-19	66%
Autonomy	14	9	1	4	10	-4	29%
Benevolence	16	8	2	5	10	-6	31%
Conformity	16	1	5	10	6	-10	63%
Enjoyment	10	8	0	2	8	-2	20%
Organizational support	12	9	3	0	12	0	0%
Tradition	11	0	1	10	1	-10	91%
Universalism	31	15	7	9	22	-9	29%
Variety	25	12	2	10	14	-11	40%
Working environment	20	12	5	3	17	-3	15%

Note. Yes = "Appropriate", Revise = "To be revised", No = "Not appropriate", $N_{final\ items}$ = Yes + Revise (CV I ≥ .70), Change = $N_{initial\ items} - N_{final\ items}$

A deep analysis of sub-themes showed that most items reaching an unsatisfactory CVI were related to specific clusters of items. For example, none of the Salary (Authority) and Modesty (Tradition) items achieved a satisfactory CVI based on predetermined cut-off values (Table 5). Other sub-themes such as Dependability (Benevolence), Enjoying research (Enjoyment), and Fairness (Organizational support) remained unchanged.

To sum up, 82 items reported a satisfactory CVI (CVI ≥ .78) shared across 10 dimensions (labeled "Appropriate" in Table 4 and 5). However, for theoretical reasons and with a view to a future further exploratory step (i.e., an exploratory factor analysis), the 40 items reporting borderline CVI (i.e., 70 ≤ CV I ≤ .78, labeled "To be revised" in Table 4 and 5) were added to the list of new items subject to further revision and adjustment.

Table 5

Summary of outcomes based on CVI per sub-theme

Dimensions	Sub-themes	$N_{initial\ items}$	Yes	Revise	No	$N_{new\ items}$	Change	No %
Ambition	Achievement	7	0	1	6	1	6	85.71%
Ambition	Career	7	3	2	2	5	2	28.57%
Ambition	Competence	8	2	4	2	6	2	25.00%
Authority	Control over resources	2	0	2	0	2	0	0.00%
Authority	Dominance over others	8	0	1	7	1	7	87.50%
Authority	Influence	5	2	2	1	4	1	20.00%
Authority	Prestige	7	1	2	4	3	4	57.14%
Authority	Salary	7	0	0	7	0	7	100.00%
Autonomy	Practical autonomy	6	2	1	3	3	3	50.00%
Autonomy	Intellectual autonomy	8	7	0	1	7	1	12.50%
Benevolence	Caring for others	7	5	1	1	6	1	14.29%
Benevolence	Dependability	1	1	0	0	1	0	0.00%
Benevolence	Relationships	8	2	1	4	3	5	62.50%
Conformity	Codes of conduct	3	1	1	1	2	1	33.33%
Conformity	Scientific norms	4	0	3	1	3	1	25.00%
Conformity	Social norms	9	0	1	8	1	8	88.89%
Enjoyment	Enjoying research	4	4	0	0	4	0	0.00%
Enjoyment	Pleasurable activities	6	4	0	2	4	2	33.33%
Organizational support	Clarity	4	3	1	0	4	0	0.00%
Organizational support	Fairness	3	3	0	0	3	0	0.00%
Organizational support	Support	5	3	2	0	5	0	0.00%
Tradition	Modesty	2	0	0	2	0	2	100.00%
Tradition	Tradition	9	0	1	8	1	8	88.89%
Universalism	Research ethics	12	10	2	0	12	0	0.00%
Universalism	Social impact	12	3	1	8	4	8	66.67%
Universalism	Sustainability	3	0	3	0	3	0	0.00%
Universalism	Tolerance	4	2	1	1	3	1	25.00%
Variety	Challenge	8	1	1	6	2	6	75.00%
Variety	Growth	4	3	1	0	4	0	0.00%
Variety	Novelty	6	5	0	1	5	1	16.67%
Variety	Variety	7	3	0	3	3	4	57.14%
Working environment	Job security & stability	5	3	1	1	4	1	20.00%
Working environment	Safety & wellbeing	5	1	2	2	3	2	40.00%
Working environment	Safety at work	10	8	2	0	10	0	0.00%

Note. Yes = "Appropriate", Revise = "To be revised", No = "Not appropriate".

Weighing items

To even out item numbers per dimension for further psychometric evaluation, the list of 122 items was further revised. Sub-themes - originally derived from the value theories underlying our item generation process to represent a broad array of value constructs - within the same dimensions were weighed against each other to calculate how many of their items were to be included (see item weighing exemplified in Table 6). To build a balanced and parsimonious instrument including an adequate number of items in terms of representativeness and reliability, we aimed to maintain a range of 8-10 items for each dimension (at least 8 each, more for dimensions with enough items to select from) and include only the top-rated items (in terms of CVI) per sub-theme accordingly.

Table 6

Item weighing exemplified

Dimensions	Sub-themes	$N_{sub\text{-}theme}$	Weighing = $N_{sub\text{-}theme}$ / $N_{dimension}$	Items to be included
Universalism	Research ethics	12	.55	6
Universalism	Social impact	4	.18	2
Universalism	Sustainability	3	.14	1
Universalism	Tolerance	3	.14	1

Qualitative assessment

Items with the same CVI were further qualitatively evaluated for their fit within the dimension by checking for overlaps in meanings with other items and their theoretical relevance. For each dimension we selected the most relevant theoretically related items covering the theoretically proposed sub-themes. Sub-themes in which no items achieved a satisfactory CVI based on predetermined cut-off values (i.e., Salary, Modesty) were eliminated. To include at least 8 items in each dimension and maintain theoretical accuracy, the Conformity and Tradition dimensions were merged together and two items scoring below .70 (CVI = .65) were included in this combined dimension.

Final item list

The final item list consisted of 97 items spread over 10 dimensions (Figure 3 and Table 7), with 10 items for all dimensions except Enjoyment (8 items) and Conformity/Tradition (9 items). In Table 7 we report the overall scale content validity index (S-CVI) for each scale. Overall, with the exemption of the Conformity/Tradition scale, all dimensions achieved a satisfactory S-CVI.

Figure 3

Academic research values dimensions

Table 7

List of items

Dimension	Sub-theme	Item	CVI	S-CVI
Ambition	Achievement	To be capable	.7	.78
	Career	To advance my career	.85	
		To get recognition for the work I do	.9	
		To win grants, scholarships, and scientific awards	.75	
		To be successful academically	.75	
	Competence	To publish in good journals	.75	
		To build my scientific reputation	.75	
		To be seen as intelligent by research peers	.75	
		To have a scientific impact	.8	
		To be visible in the scientific community	.8	
Authority	Control over resources	To have authority over research funds	.75	.77
		To have direct influence over funding decisions	.75	
	Dominance over others	To lead a research group	.75	
	Influence	To be respected as a researcher	.7	
		To be "somebody" in the research community	.75	
		To have influence in my field	.8	
		To get respect and attention for my research	.9	
	Prestige	To lead a prestigious research group	.7	
		To be invited by prestigious institutions	.7	
		To get scientific recognition	.9	
Autonomy	Freedom of action / Practical autonomy	To determine how I spend my workday	.9	.95
		To make decisions on my own	.7	
		To determine who I work with	.95	
	Freedom of thought / Intellectual autonomy	To make my own decisions about my scientific work	1	
		To be able to direct my own research	1	
		To decide my own research priorities	1	
		To define my own scientific aims	.95	
		To have a high level of professional autonomy	1	
		To try out some of my own ideas	1	
		To have freedom in choosing my research methods	1	
Benevolence	Caring for others	To prevent harm to my closest colleagues / research group	.8	.83
		To help the people in my research community	.9	
		To do work which helps colleagues	.8	
		To be supportive of colleagues	.9	
		To contribute to the flourishing of the members of the research community	.8	
		To not harm people I work with	.75	
	Dependability	To be considered a dependable and trustworthy colleague	.85	
	Relationships	To have good interactions with fellow researchers	.9	

		To be on good terms with colleagues	.75	
		To feel like I am a part of the research community	.85	
Conformity / Tradition	Scientific norms	To work with researchers who follow rules even when no one would know if they did otherwise	.65	.72
		To work in a group where researchers conform to scientific norms	.75	
		To work with researchers who respect scientific norms	.7	
		To conform to scientific norms	.75	
	Social norms	To have people within my research team get along well	.7	
	Tradition	To respect and follow well-established methodological norms	.75	
	Codes of conduct	To stay informed about changes in codes of conduct	.65	
		To resist temptation and conform to professional and scientific codes of conduct	.75	
		To work in a group where we all support the guidelines on responsible conduct of research	.8	
Enjoyment	Enjoying research	To enjoy doing research	.95	.88
		To enjoy solving scientific problems / challenges / puzzles	.9	
		To do research that makes me feel good	.85	
		To take pleasure in doing research	.85	
	Pleasurable activities	To have pleasurable experiences at work	.9	
		To enjoy my time at work	.9	
		To enjoy the company of fellow researchers	.9	
		To take pleasure in the company of interesting, smart people in the research community	.8	
Organizational support	Clarity	To have clarity about the resources provided by the university	.95	.90
		To work at a university which makes research, training, and other resources accessible	.85	
		To be clearly informed about the rules and my obligations	.8	
	Fairness	To know that the research institution distributes resources fairly	.95	
		To work at a university that administers its policies fairly	.9	
		To know that the university handles work-related processes fairly	.95	
	Support	To work for a university that assigns importance to caring for my mental and physical health	.75	
		To know that my manager/supervisor would back up the workers (with top management)	.9	
		To have a manager/supervisor who treats me well	1	
		To feel supported by the university I work at	.95	
Universalism	Research ethics	To do the work without feeling that it is morally wrong	.85	.87
		To protect scientific integrity	.85	
		To ensure honesty in my research	.9	
		To regularly verify the accuracy of my data	.95	
		To ensure that my results are replicable	.9	
		To prevent research misconduct (falsification, fabrication, and plagiarism)	.9	
	Social impact	To better the world with my research	.85	
		To make sure that my research does not have harmful consequences for others	.85	
	Sustainability	To make sure that the outcomes of my research do not have harmful consequences for nature	.75	
	Tolerance	To be willing to consider other scientific perspectives	.85	

Variety	Challenge	To experience a variety of interesting research challenges	.85	.89
	Growth	To continuously learn / develop	.95	
		To never stop learning	.9	
		To learn new skills	.9	
	Novelty	To be curious	.85	
		To explore creative ideas	.9	
		To encounter exciting new ideas	.9	
		To explore new ideas	.85	
	Variety	To do varied work	.8	
		To have cognitively stimulating experiences	.95	
Working environment	Job security and stability	To know that I will have a job in five years	.95	.88
		To have job security	1	
	Safety and wellbeing	To not be a subject of personal attacks for my research	.7	
		To have a healthy work-life balance	.85	
	Safety at work	To have a job that has good working conditions	.8	
		To work in an environment free from abusive relationships	.95	
		To work with researchers who value my mental and physical health	.85	
		To work in a safe environment	.9	
		To not be required to engage in actions I deem unsafe	.9	
		To not be required to engage in actions I deem unethical or illegal	.9	

Discussion

In this study, we use a mixed-method approach to validate a scale measuring values in academic research settings. Building on prior conceptualization and item generation work, we reduced a set of 246 academic research value items to 97, spread over 10 dimensions (Figure 3 and Table 7) based on expert input through two steps. First, aided by the grammatical, syntax, and face validity evaluation of two experts, we reduced the list to 205 items. Second, the relevance of each item was calculated through the content validity assessments of 20 experts. To our knowledge, our attempts are the first that aim to create and validate a scale measuring academic research values by integrating insights from most broadly used personal and work value measurements. In this paper, we provide detailed documentation to enhance the transparency of our methodology. This mitigates the need for future researchers to start from scratch and allows them to use our results as a springboard for further investigation. By making our findings accessible, we contribute to a growing body of knowledge that can evolve through incremental advancements rather than parallel efforts.

In some regards our results align with prior understandings of the values and motivations of researchers. Some of the best rated dimensions (based on their S-CVI) such as autonomy, benevolence, universalism, or variety overlap with the value items participants ranked as high importance in our prior study (Kis et al., 2023) as well as in other, comparable value questionnaires (Knafo & Sagiv, 2004). Items included in these dimensions referring to self determination, professional autonomy, freedom of choice, or being curious, creative, honest, and supportive are recognizable and relatable descriptors of the stereotypical or ideal researcher (Johnson & Dieckmann, 2020; Tintori, 2017) and are often presented as virtues and bases of good scientific practices (Demirutku & Güngör, 2021; English et al., 2018; Knafo & Sagiv, 2004).

The sub-themes that remained unchanged due to their items' satisfactory CVI ratings (Dependability, Enjoying research, Fairness – see Table 5) posit a similar interpretation. For the Dependability sub-theme, this could be partly due to it consisting of only one item. Still, some benevolence values have been reported to hold increased importance for researchers (Knafo & Sagiv, 2004) and being dependable and trustworthy may be considered especially relevant in a vocation geared as much towards generating dependable results as academia is. In a similar fashion, Enjoying research and solving puzzles seems to be intrinsic to being a researcher, as marked by reports on the motivations of scientists (Guerin et al., 2015; Lam, 2011). Values pertaining to Fairness and Organizational support have yet to be tested in terms of their importance to researchers. However, the growing body of literature on academics' experiences within precarious or toxic organizational environments (recent examples include: Kis et al., 2022; McKenzie, 2021; Pelletier et al., 2019; Pruit et al., 2021; Skakni et al., 2019) signals that safe, secure, and fair organizational settings are valued by researchers.

The values that are less associated with scientific work allow for an equally interesting discussion relevant for current debates about academic work. During the validation process we excluded two value sub-themes from the final scale (i.e., Salary and Modesty) based on the low content validity indices of these sub-themes. This outcome aligns with some of our prior discussions on researchers' values (Kis et al., 2023): values represented in the Salary sub-theme such as earning a high salary or increasing one's income are aligned with the theoretical underpinnings of the values literature and as such, have also been included in another scientific value scale. However, our prior results as well as that of others have consistently ranked personal income in the bottom of researchers' motivational hierarchies. Similarly, values related to tradition were ranked low in previous studies.

The exclusion of some values underscores the need for critical evaluations of the relevance and impact of each of these sub-themes. Excluding these values also emphasizes the

need to focus on those dimensions that garner broad-based support across the academic community. The absence of specific sub-themes, however, does not diminish their potential significance for academic values. The low ratings may instead suggest variability in expert opinions, as well as self-selection bias. This may be especially relevant for the Salary sub-theme. Although our invitation message noted our financial incentive, it was set up to convey our need for help and advice regarding a topic important to the scientific community (see invitation protocol in our OSF repository). Participants who chose to participate might have cared comparatively less about high salaries than some of those who did not participate. This limitation marks the need for future research to refine our validation procedures.

Approaches better suited to capture these aspects might yield benefits in terms of better understanding these now unused sub-themes which might hold substantial importance in academic settings. The tension between a normative versus descriptive approach to values in academic research illustrates one of our study's limitations. While invaluable, the input from our experts may carry inherent biases towards normative values, potentially skewing the scale towards idealized perceptions rather than describing empirical realities. Acknowledging these biases is crucial as it informs the interpretation of our results and guides future research to address these gaps.

Another limitation of our methods stems from some of our decisions. We made efforts to minimize inconsistencies and incorporate empirical evidence based on best practices, but – partly due to a lack of straightforward guidelines and the complexity of our construct – our process necessarily incorporated some subjective choices. For example, qualitative stages involving coding and excluding items in both Step 1 and 2 could have yielded different outcomes for different researchers. Arguably, many (if not most) of these subjective choices were unavoidable due to the nature of our approach. We also tried to provide detailed documentation of each of our steps to ensure full transparency.

Moreover, the sample selection, although adequate for initial explorations and set up to encounter uncontrollable variety in responses, limited our scope to collecting the opinions of researchers working in the Netherlands. This sample is unlikely to fully capture the diversity of academic values across different disciplines and cultural contexts. As such, it should be interpreted responsibly in terms of the generalizability of our current findings. The sample size also affected our ability to conduct measurement invariance tests across disciplines and contexts, which are key to valid and generalizable measures (Hussey & Hughes, 2020). The reliance on expert judgments in the validation process might overlook nuances that could be captured through broader community engagement or more varied methodological approaches. For example, Mason et al. (2023) illustrate how non-expert perceptions of psychological measures can inform and enrich the validation process via qualitative methods, reinforcing the importance of integrating various populations (and approaches) to fully capture and validate new measures. Testing generalizability and extending our inquiries to more diverse groups will be a goal for future research.

Several of these limitations may be addressed in subsequent stages of the scale development process, since the scale still needs to be further evaluated before being reliably applicable. An important future development step is to involve larger and more diverse samples of researchers to ensure a more representative range of the target population. Additionally, methodological steps will include examining the factorial structure of the instrument through exploratory methods (i.e., Exploratory Factor Analysis) followed by confirmatory methods (i.e., Confirmatory Factor Analysis), and evaluating the external validity of the scale (i.e., the relationship of the emerged factors with external measures).

In conclusion, this research advances the field by developing a new measure of academic values and critically examines the complexities involved in such a task. The implications of our findings extend beyond scale development. Some of our results provide

insights into the normative structures that influence academic behaviors and priorities. Others are relevant in terms of detailing the methodological steps and choices involved in content validation of psychological scales.

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Brooks Cole.

Almanasreh, E., Moles, R., & Chen, T. F. (2019). Evaluation of methods used for estimating content validity. *Research in Social and Administrative Pharmacy*, *15*(2), 214–221. https://doi.org/10.1016/j.sapharm.2018.03.066

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *The Standards for Educational and Psychological Testing*. https://www.apa.org/science/programs/testing/standards

Anastasi, A. (1950). The concept of validity in the interpretation of test scores. *Educational and Psychological Measurement*, *10*, 67–78. https://doi.org/10.1177/001316445001000105

Anastasi, A. (1988). *Psychological testing* (6th ed., pp. xiv, 817). Macmillan Publishing Co, Inc.

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer. *Frontiers in Public Health*, *6*. https://www.frontiersin.org/articles/10.3389/fpubh.2018.00149

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brown, T. (2010). Construct Validity: A Unitary Concept for Occupational Therapy Assessment and Measurement. *Hong Kong Journal of Occupational Therapy*, *20*(1), 30–42. https://doi.org/10.1016/S1569-1861(10)70056-5

Delgado-Rico, E., Carrctero-Dios, H., & Ruch, W. (2012). Content validity evidences in test development: An applied perspective. *International Journal of Clinical and Health Psychology*, *12*(3), 449–459.

Demirutku, K., & Güngör, E. (2021). Content and Structure of Scientific Values. *Psikoloji Çalışmaları / Studies in Psychology*, *41*(2), 459–489. https://doi.org/10.26650/SP2019-0142

Douglas, H. (2023). The importance of values for science. *Interdisciplinary Science Reviews*, *48*(2), 251–263. https://doi.org/10.1080/03080188.2023.2191559

English, T., Antes, A. L., Baldwin, K. A., & DuBois, J. M. (2018). Development and Preliminary Validation of a New Measure of Values in Scientific Work. *Science and Engineering Ethics*, *24*(2), 393–418. https://doi.org/10.1007/s11948-017-9896-0

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct Validation in Social and Personality Research: Current Practice and Recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378.

Gable, R. K., & Wolf, M. B. (1993). *Instrument Development in the Affective Domain*. Springer Netherlands. https://doi.org/10.1007/978-94-011-1400-4

Grant, J. S., & Davis, L. L. (1997). Selection and use of content experts for instrument development. *Research in Nursing & Health*, *20*(3), 269–274. https://doi.org/10.1002/(sici)1098-240x(199706)20:3<269::aid-nur9>3.0.co;2-g

Guerin, C., Jayatilaka, A., & Ranasinghe, D. (2015). Why start a higher degree by research? An exploratory factor analysis of motivations to undertake doctoral studies. *Higher Education Research & Development*, *34*(1), 89–104. https://doi.org/10.1080/07294360.2014.934663

Hardesty, D. M., & Bearden, W. O. (2004). The use of expert judges in scale development: Implications for improving face validity of measures of unobservable constructs. *Journal of Business Research*, *57*(2), 98–107.

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment*, *7*(3), 238–247. https://doi.org/10.1037/1040-3590.7.3.238

Hussey, I., & Hughes, S. (2020). Hidden Invalidity Among 15 Commonly Used Measures in Social and Personality Psychology. *Advances in Methods and Practices in Psychological Science*, *3*(2), 166–184.

Johnson, B. B., & Dieckmann, N. F. (2020). Americans' views of scientists' motivations for scientific work. *Public Understanding of Science*, *29*(1), 2–20. https://doi.org/10.1177/0963662519880319

Kis, A., Tur, E. M., Lakens, D., Vaesen, K., & Houkes, W. (2022). Leaving academia: PhD attrition and unhealthy research environments. *PLOS ONE*, *17*(10), e0274976. https://doi.org/10.1371/journal.pone.0274976

Kis, A., Tur, E. M., Vaesen, K., Houkes, W., & Lakens, D. (2023). *Academic Research Values: Conceptualization and Initial Steps of Measure Development*. PsyArXiv. https://doi.org/10.31234/osf.io/qzkew

Knafo, A., & Sagiv, L. (2004). Values and work environment: Mapping 32 occupations. *European Journal of Psychology of Education*, *19*(3), 255–273. https://doi.org/10.1007/BF03173223

Koller, I., Levenson, M. R., & Glück, J. (2017). What Do You Think You Are Measuring? A Mixed-Methods Procedure for Assessing the Content Validity of Test Items and Theory-Based Scaling. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00126

Lam, A. (2011). What motivates academic scientists to engage in research commercialization: 'Gold', 'ribbon' or 'puzzle'? *Research Policy*, *40*(10), 1354–1368. https://doi.org/10.1016/j.respol.2011.09.002

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, *28*(4), 563–575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Lynn, M. R. (1986). Determination and Quantification Of Content Validity. *Nursing Research*, *35*(6), 382.

Mason, J., Pownall, M., Palmer, A., & Azevedo, F. (2023). Investigating Lay Perceptions of Psychological Measures: A Registered Report. *Social Psychological Bulletin*, *18*, 1–32. https://doi.org/10.32872/spb.9383

McKenzie, L. (2021). Unequal expressions: Emotions and narratives of leaving and remaining in precarious academia. *Social Anthropology*, *29*(2), 527–542. https://doi.org/10.1111/1469-8676.13011

Messick, S. (1989). Meaning and Values in Test Validation: The Science and Ethics of Assessment. *Educational Researcher*, *18*(2), 5–11. https://doi.org/10.2307/1175249

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Morgado, F. F. R., Meireles, J. F. F., Neves, C. M., Amaral, A. C. S., & Ferreira, M. E. C. (2017). Scale development: Ten main limitations and recommendations to improve future research practices. *Psicologia: Reflexão e Crítica*, *30*(1), 3. https://doi.org/10.1186/s41155-016-0057-1

Nevo, B. (1985). Face validity revisited. *Journal of Educational Measurement*, *22*(4), 287–293. https://doi.org/10.1111/j.1745-3984.1985.tb01065.x

Pelletier, K. L., Kottke, J. L., & Sirotnik, B. W. (2019). The toxic triangle in academia: A case analysis of the emergence and manifestation of toxicity in a public university. *Leadership*, *15*(4), 405–432. https://doi.org/10.1177/1742715018773828

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health*, *29*(5), 489–497. https://doi.org/10.1002/nur.20147

Pruit, J., Pruit, A., & Rambo, C. (2021). "Suck It up, Buttercup": Status Silencing and the Maintenance of Toxic Masculinity in Academia. *Studies in Symbolic Interaction*, *52*, 95–114. https://doi.org/10.1108/S0163-239620210000052007

Roebianto, Roebianto, Savitri, Aulia, Suciyana, & Mubarokah. (2023). Content validity: Definition and procedure of content validation in psychological research. *Testing, Psychometrics, Methodology in Applied Psychology*, *30*(1), 5–18. https://doi.org/10.4473/TPM30.1.1

Rubio, D. M., Berg-Weger, M., Tebb, S. S., Lee, E. S., & Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in social work research. *Social Work Research*, *27*(2), 94–104. https://doi.org/10.1093/swr/27.2.94

Sagiv, L., & Roccas, S. (2021). How Do Values Affect Behavior? Let Me Count the Ways. *Personality and Social Psychology Review*, 108886832110159. https://doi.org/10.1177/10888683211015975

Sagiv, L., Roccas, S., Cieciuch, J., & Schwartz, S. H. (2017). Personal values in human life. *Nature Human Behaviour*, *1*(9), 630–639. https://doi.org/10.1038/s41562-017-0185-3

Schwartz, S. H. (1992). Universals in the Content and Structure of Values: Theoretical Advances and Empirical Tests in 20 Countries. In *Advances in Experimental Social Psychology* (Vol. 25, pp. 1–65). Elsevier. https://doi.org/10.1016/S0065-2601(08)60281-6

Schwartz, S. H., Cieciuch, J., Vecchione, M., Davidov, E., Fischer, R., Beierlein, C., Ramos, A., Verkasalo, M., Lönnqvist, J.-E., Demirutku, K., Dirilen-Gumus, O., & Konty, M. (2012). Refining the theory of basic individual values. *Journal of Personality and Social Psychology*, *103*(4), 663–688. https://doi.org/10.1037/a0029393

Shrotryia, V. K., & Dhanda, U. (2019). Content Validity of Assessment Instrument for Employee Engagement. *Sage Open*, *9*(1). https://journals.sagepub.com/doi/10.1177/2158244018821751

Skakni, I., Calatrava Moreno, M. del C., Seuba, M. C., & McAlpine, L. (2019). Hanging tough: Post-PhD researchers dealing with career uncertainty. *Higher Education Research & Development*, *38*(7), 1489–1503. https://doi.org/10.1080/07294360.2019.1657806

Spoto, A., Nucci, M., Prunetti, E., & Vicovaro, M. (2023). Improving Content Validity Evaluation of Assessment Instruments Through Formal Content Validity Analysis. *Psychological Methods*. https://doi.org/10.1037/met0000545

Tintori, A. (2017). The most common stereotypes about science and scientists: What scholars know. In *Turn on the light on science: A research-based guide to break down popular stereotypes about science and scientists*. Ubiquity Press. https://www.ubiquitypress.com/site/chapters/m/10.5334/bba.b/

UNESCO. (2021). *UNESCO Recommendation on Open Science* (p. 34). UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000379949/PDF/379949eng.pdf.multi

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: The kappa statistic. *Family Medicine*, *37*(5), 360–363.

Zamanzadeh, V., Ghahramanian, A., Rassouli, M., Abbaszadeh, A., Alavi-Majd, H., & Nikanfar, A.-R. (2015). Design and Implementation Content Validity Study:

Development of an instrument for measuring Patient-Centered Communication. *Journal of Caring Sciences*, *4*(2), 165–178. https://doi.org/10.15171/jcs.2015.017