Frontline Learning Research
An official journal of EARLI

Frontline **L**earning **R**esearch Vol.8 No. 6 (2020) 1 - 37
ISSN 2295-3159

Ensuring content validity of psychological and educational tests – the role of experts

Klaus Beck [a]

[a]Johannes Gutenberg-University Mainz, Germany

Article received 18 June 2019 / Article revised 29 July 2020 / Accepted 14 August / Available online 4 September

Abstract

Many test developers try to ensure the content validity of their tests by having external experts review the items, e.g. in terms of relevance, difficulty, or clarity. Although this approach is widely accepted, a closer look reveals several pitfalls need to be avoided if experts' advice is to be truly helpful. The purpose of this paper is to exemplarily describe and critically analyse widespread practices of involving experts to ensure the content validity of tests. First, I offer a classification of tasks that experts are given by test developers, as reported on in the respective literature. Second, taking an exploratory approach by means of a qualitative meta-analysis, I review a sample of reports on test development (N = 72) to identify the common current procedures for selecting and consulting experts. Results indicate that often the choice of experts seems to be somewhat arbitrary, the questions posed to experts lack precision, and the methods used to evaluate experts' feedback are questionable. Third, given these findings I explore in more depth what prerequisites are necessary for their contributions to be useful in ensuring the content validity of tests. Main results are (i) that test developers, contrary to some practice, should not ask for information that they can reliably ascertain themselves ("truth-apt statements"). (ii) Average values from the answers of the experts which are often calculated rarely provide reliable information about the quality of test items or a test. (iii) Making judgements about some aspects of the quality of test items (e.g. comprehensibility, plausibility of distractors) could lead experts to unreliable speculations. (iv)When several experts respond by giving incompatible or even contradictory answers, there is almost always no criterion enabling a decision between them. In conclusion, explicit guidelines on this matter need to be elaborated and standardised (above all, by the AERA, APA, and NCME "Standards").

Keywords: test development; content validity; expert; qualitative meta-analysis

Corresponding author: Klaus Beck, Johannes Gutenberg-University Mainz, Faculty of Law, Management and Economics, Department of Business Education, 55099 Mainz, Germany, beck@uni-mainz.de
DOI: *https://doi.org/10.14786/flr.v8i6.517*

"If all experts are united, caution should be exercised."
Albert Einstein

1. Introduction

In the latest edition of the Standards for Educational and Psychological Testing (AERA, APA & NCME, 2014) the multi-dimensionality of test validity is stressed once again. Validity is considered not merely an attribute of a single test or a given assessment procedure, but rather of the interpretation of test scores[1], be it in the context of diagnosing individuals, identifying group properties, determining causal relationships between latent traits, or developing practical educational measures. When taking this approach to validity several facets emerge which might affect the accuracy of conclusions drawn from a measurement result.

The issue of validity has been explored extensively in the literature. In this paper, one particular facet of validity essential to adequate test score interpretations is explored: content validity.[2] A basic tenet of test development is that a test instrument should measure what it claims to measure. Therefore, the content of such instruments has to reflect or correspond as accurately as possible to the real-world issues it is intended to assess. If this condition is not satisfied, interpretation of the outcomes of an assessment will be inappropriate.

The difficulty begins with defining what is to be measured, especially in regard to latent attributes, which is usual in the fields of psychology and education. Following Boring (1923) some psychologists have taken the view that the latent object measured is defined by what the assessment instrument measures—a variant of early behaviourism also held in a certain sense by psychological constructionists in making use of operational definitions (e.g., Van der Maas, Kan, & Borsboom, 2014). With good reason it was argued early on in seminal debates that this interpretation is not acceptable (e.g., Miles, 1957). Of course, test developers have at least a rough idea of the object they want to measure when creating a new test. That idea represents or designates the content in relation to which validity only can be examined. The clearer this idea, the more precisely content validity can be investigated.[3] And the greater the validity of test content being assessed, the lower the risk of inadequate interpretations of test scores (given the satisfactory quality of all other properties that a test should exhibit). Thus, content validity is an essential attribute of all psychological measurement instruments (Lynn, 1986), and understanding how it can be achieved is of paramount importance. For this reason, test developers are interested in enhancing the content validity of their instruments. It is widely accepted as evidence of having reached this goal if experts agree that this claim has been met.

In Section 2, I give a brief overview of the current practices of involving experts in the development of tests. Here I explore the rationale for involving experts in the process of ensuring content validity of tests or test items. Additionally, I describe the cognitive processes that experts have to go through when they do their job. In Section 3, I offer a categorisation schema for the numerous tasks test developers ask experts to perform. I conduct a qualitative meta-analysis of a sample of 72 reports on test development and present a summary of the procedures test developers took when consulting experts to ensure the validity of the content of test items or tests. In Section 4, I describe the wide variety of practices currently employed and explain why they are often unreliable. In Section 5, I discuss some of

[1] In the 1985 edition of the "Standards" it says: "The inferences regarding specific uses of a test are validated, not the test itself" (p. 9).
[2] Though being judged as technically incorrect and therefore theoretically not helpful in the early years after inception, the concept of content validity has survived and still enjoys attention in the relevant literature (Guion, 1977; Sireci, 1998).
[3] The first author to develop a quantification of content validity is presumably Lawshe (1975). Several modifications of his measure have been developed (Wilson, Pan, & Schumsky, 2012).

the main problems encountered when consulting experts on test development. In Section 6, I draw conclusions from the main findings of the data analyses and specify the need for improvement of current practices in consulting experts on test development.

2. Experts' advice on how to ensure content validity: theoretical background

When developing psychological or educational tests the typical way to ensure content validity is to make use of the expertise[4] of professionals in the field (Allen & Yen, 2002; Reynolds, Livingston, & Willson, 2009). By a priori assumption, experts can determine whether or not the content of a test is "sensitive" to variation in the formation of a latent attribute that the test or one of its items is directed toward. Due to the innumerable situations a person may find him or herself in, this latent attribute could invoke an infinite number of perceptions, thoughts, reactions, or actions. In this sense, a group of test items is understood as a sample of situational constellations to which a person might react (or on which a person might act) relatively adequately (Kerlinger, 1986; Anastasia & Urbina, 1997).

In terms of ensuring the content validity of a test, an expert is asked to judge, assess, or rate the extent to which the items of a test adequately represent the infinite number of situational constellations by which the latent attribute to be measured stimulates a certain behaviour in test takers. This implies that the expert needs to do the following:

- identify an understanding of the latent attribute to be measured;
- determine the universe of possible situational constellations by which the latent attribute is activated;
- consider all the behavioural patterns being caused by the latent attribute in dealing with these situational constellations;
- relate these constellations to the test and its items and to the respective latent attribute to be stimulated, that is, the construct, and examine whether these relationships are adequate in terms of the aim(s) of the test;
- determine whether the test items will function as adequate stimulators for different groups of test takers (e.g., in terms of language, level of education, social background, etc.);
- provide feedback to the test developer(s) in a way that allows them to make decisions concerning the validity of the test items (Grant & Davis, 1997, p. 272, col. 1-2). Experts can give feedback informally, for example during structured interviews, or formally, for example in written reports or surveys with "yes" or "no" ratings of acceptability or ratings on Likert scales. Anderson et al. (2015, p. 22), for example, recommend following an online Likert-based procedure when working with large data banks and large pools of experts, and using latent trait models "to control for between-rater severity, evaluate interrater consistency, and provide item-level diagnostic statistics."

Usually, several experts are consulted to evaluate content validity. If experts are scattered around a country (e.g. across the US) they might have to convene in person to discuss their ratings, if different, until consensus on the (good or bad) quality of an item or a test is reached or a majority vote

[4] In this paper the meaning of the terms *expert* and *expertise* are not as specialised as in Ericsson and Smith (1991). Rather, the terms are used more generally to convey the meaning of *comprehensive and authoritative knowledge of or skill in a particular area*.

is taken. The greater the number of experts consulted, the more likely it is that they do not meet in person, and that quantifying procedures are used to obtain information from them (Lawshe, 1975; Thorn & Deitz, 1989). As a qualitative alternative, Delphi studies may be conducted in which experts state their opinions in written form, then the test developer bundles those opinions and distributes them repeatedly until consensus, or at least a majority consensus, is reached (see e.g., Lohse-Bossenz, Kunina-Habenicht, & Kunter, 2013; Messmer & Brea, 2015).

Test developers also often assign different tasks to experts. For example, they pose different questions to the experts, sometimes asking one group to judge all items as a whole and the other group to judge each item separately, or sometimes by recruiting multiple groups of experts and asking each group to evaluate one validity-related aspect of the items only (see e.g., Aydın & Uzuntiryaki, 2009, p. 871; Mesmer-Magnus et al., 2010, pp. 514-515; Jenßen, Dunekacke, & Blömeke, 2015, pp. 21-22; cf. Appendix 2, no. 6). Such aspects include the representativeness, clarity, relevance, conciseness, or technical adequacy of single items or the test as a whole (see Table 1). When consulting more than one expert, the question arises as to how to deal with different or even incommensurable judgements rendered by them. If the experts' feedback is qualitative, the test developer may have to grapple with diverging arguments, weigh them, and finally decide which are the most plausible. Here, the question arises as to what quality the arguments should have in order to override the opinion of at least some experts, and why the test developer should ask experts at all if in the end he or she will be making the final decision him- or herself. If the experts' feedback is quantitative (e.g., ratings on Likert scales), usually an arithmetic mean of these ratings along with its standard deviation is computed and then checked to determine whether certain meaningful thresholds in these values have been exceeded or underrun. Again, the test developer has to make a decision, namely, to set the boundaries of the range within which, in the end, an arithmetic mean or a standard deviation shall be acceptable (see e.g., Jenßen, Dunekacke, & Blömeke, 2015, pp. 22-23; cf. Appendix 2, no. 6). Again, the question arises as to which arguments justify the determination of those boundaries.

The aforementioned procedures are followed in an attempt to reach at least broad consensus among external experts, and also between those experts and the test developer(s) on the adequacy of test items. Such consensus tends to be used as a justification for the inclusion of the items and as an argument for the overall quality of the test (even though it is based on the implicit fallible, if not false, assumption that this is a valid and authoritative indicator of the appropriateness of the content of a test).

These approaches also seem to be fully in line with the rationale for ensuring validity by "evidence based on test content" declared by the Standards Management Committee of AERA, APA, and NCME (2014, pp. 14-15), which exemplify the role of experts by four functions:

- assigning test items to content-specific categories or facets of an occupation (p. 14, col. 2);
- judging "the representativeness of the chosen set of items" (p. 14, col. 2);
- rating the "relative importance, criticality and/or frequency" (p. 14, col. 2) of job observation-based item pools; and
- avoiding construct-irrelevant sources of variance caused by inadequate wording of items (p. 15, col. 1; see also Standards 1.9, p. 25 and 4.8, p. 88.).

It is unclear whether these four functions are meant to be exclusive. Presumably, this is not the case because there is no mention of why experts should not be allowed to perform more than these four kinds of tasks related to ensuring content validity. Messick stated that experts' "judgements should focus on such features as readability level, freedom from ambiguity and irrelevancy, appropriateness of keyed answers and distractors, relevance and demand characteristics of the task format, and clarity of instructions" (Messick, 1987, p. 55). Similarly, Kane (2013a, p. 5), citing Angoff, pointed out (and did not criticise) that "(m)ost of the early tests of mental ability and many current standardised tests of various kinds … have been justified primarily in terms of "a review of the test content by subject matter experts" (Angoff, 1988, p. 22). Elsewhere, Kane stressed that he did not claim "that the blessing of an

expert committee is, in itself, adequate for the validation of an achievement-based [test; K.B.] interpretation, but [he; K.B.] would expect an achievement test to pass this kind of challenge" (Kane 2013b, p. 121).

Though the meaning of validity has changed over time (Baker, 2013; Shepard, 2013) and now is derived from an argument-based approach, the facets of validity as previously discussed have not disappeared. Their status has only been altered by now being subordinated to the modern understanding of validity as a problem of quality of "interpretation and use of test scores" (Kane, 2013a, p. 2).[5]

It is worth pointing out that test developers draw on experts to ensure not only content validity but also construct validity (factorial validity, convergent validity, and discriminant validity), concurrent validity, cognitive validity[6] or other variations of validity.[7] In addition, experts are consulted to judge or to give advice with regard to statistical aspects of test development. Moreover, they are needed to assess the adequacy of test takers' answers to open-ended questions on a test.[8] Experts' additional opinions are in demand as assessments in education, at least in the United States, seem to shift from measurement tools to policy levers establishing test-based accountability policies, a development which launches new test functions and is considered to be a matter of consequential validity (Henig, 2013; Shepard, 2013; Welner, 2013; AERA, APA, & NCME, 2014, p. 14). In this paper only content validity in the sense described above is explored.

3. Types of information needed from experts

Test developers do not always consult experts to ensure content validity. In some cases, they feel confident that they are able to fulfil this quality criterion without help because they consider themselves to be experts. Further, they might not know or have contact with the "very best" experts who are willing and able to get involved in the development of their test(s). However, if they decide to consult experts, they may have several questions they want to pose to them. In Table 1 is a list of 31 types of tasks that experts might be asked to perform.

The list of types of tasks for experts was developed in three successive stages. First, ten studies were evaluated in terms of the tasks experts were asked to perform. These tasks were grouped under general headings and according to the stages of the test development process in which they were performed. Then, as each subsequent study was analysed, any new task for experts found was placed under one of the general headings if possible or was put into a new category with a new heading. In the end, the categories were reviewed and any possible task for experts I felt, from my own experience with test development, was still missing from the list was added. Overall, 31 types of tasks for experts were

[5] To be more exact, Kane emphasises that this terminology gives equal weight to *interpretation* and *use*, as in his formulation "interpretation/use argument" (IUA) (Kane, 2013a, p. 2).

[6] Though related to content validity, the construct *cognitive validity* points in particular to mental processes of test takers, while content validity is devoted mainly to domain-specific aspects of wording and the purport of test items. As a matter of fact, "cognitive validity" in particular designates the problem of whether an item elicits thought processes of adequate complexity (Field, 2013; Smith, 2017). Nevertheless, the relationship between the two constructs needs to be clarified.

[7] With regard to types of validity there are many more circulating in the literature. See e.g., Messick (1990). Newton and Shaw (2014, pp. 7-8) list 151 "(k)inds of validity that have been proposed over the decades" (p. 8, Table 1.3).

[8] This task often is done by *raters*. In a sense they also are experts. However, their function is clearly distinct from that of experts involved in test development. Nevertheless, White's (2018) sophisticated discussion on performance standards for raters could be transferred and applied in part to the issue of experts involved in ensuring the content validity of tests.

identified. Thus, this inductive and deductive approach to identifying tasks for experts is not intended to be understood as an exhaustive list.

Table 1

Experts' contributions to ensure content validity

No.	Tasks *(in the general order of needs arising during the test development process)*	Type of expert input*
1	defining/delimiting the respective content area/domain	subjective (S)
2	identifying domain specificity of items/tasks/problems	subjective (S)
3	assessing the accuracy and appropriateness of translation/transfer of test items from one language/cultural context to another	subjective (S)
4	judging/ranking items/tasks/problems along domain-/work-related criteria: a) relevance (e.g., as facet of a complex competence) …….. b) importance/significance (e.g., for a certain profession) …..... c) representativeness/typicality (e.g., in a field of activity) …….... d) dangerousness (risk of harm or endangering others or damaging materials in completing a test item) ……..... e) occurrence on the job (yes/no) …….... f) frequency of occurrences on the job ……..... g) content dimensions of the respective domain (e.g., subdimensions such as marketing, finance, accounting in business; sale, service, repair in car workshops; emergency aid, pharmaceuticals in health care, etc.)	 subjective (S) subjective (S) subjective (S) subjective (S) truth-apt $_{objective}$ (T_o) truth-apt $_{unknown}$ (T_u)* truth-apt $_{objectively}$ (T_o)
5	judging/ranking/assigning items/tasks/problems according to … a) test taker related criteria aa) difficulty/complexity (e.g., for a certain group of test takers) …… ab) gender-specificity/-neutrality ..…… ac) cultural fairness …… ad) aspiration level (high, medium, low; by scale) …… b) psychological categories (e.g., classifying items into stages of cognitive, affective, psycho-motoric taxonomies)	 subjective (S) truth-apt $_{unknown}$ (T_u)* subjective (S) subjective (S) T_o
6	evaluating the representativeness of a number of items as a sample drawn from the universe of a particular domain	truth-apt $_{unknown}$ (T_u)*
7	ensuring the quality of wording/phrasing of items in terms of … a) clarity (e.g., of phrasing) ……..…….... b) understandability (e.g., foreign/abstract words) …….. c) grammar (correctness) ……...…… d) unambiguousness (uniqueness) …….... e) consistency (e.g., in use of terms) ……...……..…….. *in the case of MC format:* f) correctness of attractors ……..…….. g) falsity of distractors …….. h) plausibility of distractors ……..	 subjective (S) subjective (S) truth-apt $_{objectively}$ (T_o) subjective (S) truth-apt $_{objectively}$ (T_o) truth-apt $_{objectively}$ (T_o) truth-apt $_{objectively}$ (T_o) subjective (S)
8	categorising items according to their … a) feasibility (e.g., time onstraints)……..……... b) suitability/appropriateness (e.g., to be presented in a dynamic version) ….. c) curricular sequence in learning/educational process …….... d) logical sequence (e.g., not presupposing the solution of a following item) .. e) psychological sequence (e.g., from easy/simple to difficult/complex) …..... f) curricular significance (high, medium, low; by scale) …….. g) opportunities to have been learned …….....	 subjective (S) subjective (S) truth-apt $_{objectively}$ (T_o) truth-apt $_{objectively}$ (T_o) subjective (S) subjective (S) truth-apt $_{unknown}$ (T_u)*

Note. * T_u: truth-apt, but unknown: no objective knowledge available yet.

To analyse procedural practices, differentiation is made provisionally among three types of information[9] test developers need. The *first type* is truth-apt[10] in the analytical sense of propositional logic. This means that a statement is capable of being true or false and that it is objectively (i.e., intersubjectively) testable whether this statement is in fact true or false (indicated in Table 1 as T_o). For example, experts may be asked whether an attractor on a multiple choice test is correct, the answer being either true or not true (e.g., 7f). The *second type* is truth-apt as well, but it states information which is not yet known (indicated as T_u). For example, if asked how many traffic accidents occurred on a particular holiday throughout the world, or how often a certain cognitive performance has to be shown by an apprentice per work day, there will be a true answer but it is not available (is unknown) as long as this matter has not been researched (e.g. 4f). The *third type* asks for an expert's opinion (i.e., a valuation, estimation, or perception) regarding, for example, the relevance or understandability or feasibility of a test item or an entire test (see e.g., 4a, 7b, 8a). Such statements appear as individual judgements and therefore are inherently subjective (indicated with S). They may differ from person to person without possible rebuttal, meaning one can only agree or disagree with them. Statements of this type cannot be "true" or "false", and it would not make sense to attribute the property "truth-apt" to them. Likewise, it would not make sense to take a position on a "true" statement by saying that one does not like it or that one is ready to agree with it. As shown below, these distinctions are of fundamental importance when dealing with questions to, and answers from, experts.

4. Procedures for ensuring content validity

To gain insight into the procedures adopted by test developers to ensure content validity by a qualitative meta-analysis I reviewed a sample of 72 published reports. In Germany between 2009 and 2015 two large research programmes funded by the Ministry of Education and Research were launched to promote the development of modelling and measuring competences and skills in the fields of academic education (KoKoHs) and vocational education (ASCOT). In KoKoHs and its thematic context 24 project groups have been researching the measurement of general and domain-specific academic competences (Pant et al., 2016). The KoKoHs programme received roughly €13 million in funding.[11] In ASCOT and its thematic context six project groups have been developing mainly IT-based test instruments in the fields of commerce, mechanics, and health care for use in vocational education (Beck, Landenberger, & Oser, 2016). This initiative received approximately €7 million in funding.[12] Initial findings of KoKoHs and ASCOT were published in 2014 and 2015. Both programmes were treated as the first sections of a far-reaching research project, and were exclusively devoted to test development.[13] Within that period findings from five other projects on test development conducted in Germany (funded by other entities) were published and have therefore been included in the sample.

[9] The reason for this simplification is that, as a rule, reports on test development do not explore this issue in depth or even touch upon it. More specific differentiations are discussed below (Section 5).

[10] There is a longstanding and endless philosophical discussion on the notion of and theories surrounding *truth*. In the present context truth-aptness is understood as the meaning argued by correspondence theorists in the framework of Critical Rationalism in the sense of Popper, or Analytic Philosophy in the sense of Quine; see e.g., Jackson, Oppy, & Smith, 1994; Dodd, 2002.

[11] Around 220 researchers involved in roughly 70 single projects conducted in 12 federal states of Germany. Detailed information on measuring instruments developed in KoKoHs is given by Zlatkin-Troitschanskaia et al. 2020. See also: https://www.kompetenzen-im-hochschulsektor.de/kokohs-2011-2015/.

[12] Including more than 12,000 test takers at approximately 300 schools in 13 federal states of Germany. For more information on ASCOT, see: https://www.bmbf.de/pub/Berufsbildungsbericht_2016_eng.pdf.

[13] In the meanwhile, the follow-up phases have started being dedicated to questions of application and implementation (KoKoHs II [https://www.blogs.uni-mainz.de/fb03-kokohs-eng/] and ASCOT+ [https://www.ascot-vet.net/de/forschungs-und-transferinitiative-ascot.html]).

I also accessed two relevant American journals: 'Educational and Psychological Measurement' and 'Measurement and Evaluation in Counselling and Development'. Unlike many other journals in the area of diagnostics, they offer a special section for reports on test development in various subject areas, thus allowing easy access to the material of interest. I explored volumes from 2009 to 2015 of both journals, the same period during which the projects in Germany were being conducted.

Thus, sampling was convenient and random: "convenient" with respect to the sources from which the test development reports were drawn and "random" for the single reports which were funded independently of each other within large research programmes (Germany) or published independently of each other in the two selected journals (US). As the aim of this study is not to present representative data on its subject, but rather to give an exploratory overview of the range of procedures dealing with the inclusion of experts in ensuring content validity of tests or test items, this approach to sampling does not result in a relevant bias.

The overall sample included 72 reports on test development,[14] which I analysed as to how content validity is treated, whether experts were involved, and if so, how the experts were selected, how many of them were consulted, and what they were asked to contribute to ensure content validity. From this perspective the test content (whether psychological or educational, whether focusing on teachers, students, or apprentices), national context, and publication medium (journal or book chapter) were irrelevant, as the procedures to ensure content validity can be described and compared independently of these conditions. [15]

Of the 72 reports, 16 did not rely on experts to ensure content validity (see Appendix 1, col. 3(a): "n"): In two of them, experts' advice was purposefully avoided because the researchers felt capable of assessing the quality of their test (Appendix 1, no. 53, 72), and in the others[16] no mention was made of the use of external experts.

In the remaining 56 reports (78% of the sample) some information was provided concerning the use of experts and their functions in the process of ensuring content validity. In 10 of them[17] experts were consulted but no clear description was given about the task(s) they performed. Overall, none of the reports in the sample provided an adequate description of the criteria used to determine the exact role and contribution of experts in ensuring content validity (cf. Appendix 1 col. 3 to 6). Only three of the reports[18] came close to doing so.

In 18 reports only one type of information was requested from experts (cf. Appendix 1, col. 7) whereas in others several types of information were sought (e.g., Appendix 1, no. 25 and 35): between seven and eight questions were posed to each expert.[19] Overall, experts were asked 113 questions related to all types of information relevant to content validity (cf. Appendix 3) over the course of all test development projects in the sample. Their support was requested mostly for items concerning domain-specific characteristics (cf. Appendix 3, col. 4, type no. 4: 41 of all 113 requests) and quality of wording (type no. 7: 33 requests). Some information types listed in Table 1 did not seem to be needed very often from external experts: Type no. 1: 'definition of a content area' was requested twice only (cf.

[14] KoKoHs: 18 reports; ASCOT: 4 reports; other projects conducted in Germany: 5 reports; *Educational & Psychological Measurement*: 22 reports; and *Measurement & Evaluation* in *Counselling & Development*: 23 reports.

[15] This holds also for the selection of experts in the German programmes KoKoHs and ASCOT. Though different in regard to their research topics, they both had to deal with the question of content validity.

[16] Appendix 1, no. 3, 16, 23, 24, 26, 30, 38, 46, 48, 49, 51, 56, 59, 60.

[17] See Appendix 1, no. 8, 12, 15, 41, 42, 47, 65 – 68.

[18] See Appendix 1, col. 6, no. 17 (missing only a qualitative report on the results of consultation with experts), no. 50 (missing a quantitative report) and col. 3(c), no. 55 (missing information on criteria and method(s) for selecting experts).

[19] In the 46 projects reporting on the questions posed to experts, on average two or three questions were asked.

Appendix 3, col. 4), type no. 6: 'representativeness of an item sample' was requested five times. Other types of relevant information were not requested at all.[20]

Across all projects in which the type of questions posed to experts was reported, the majority of the experts' responses fell into the category "subjective" that is requiring approval or disapproval (Table 2).[21]

Table 2

Distribution of experts' responses to all types of questions posed

Truth-apt "objective" T_o	**Truth-apt "unknown" T_u**	**Requiring approval or disapproval "subjective" S**
12	5	96

Of the 56 reports that included the involvement of experts, 40 provided information on the number of experts they contacted. In four other reports experts were said to have been consulted but the number was not stated. In two other reports only the number of one subgroup of experts was stated. Overall, at least 1,261 experts plus an undisclosed additional number are claimed to have been involved in the 46 projects in which a number of experts consulted was stated. This means that test developers asked an average of at least 27 experts for help to ensure the content validity of their instrument.[22] It is not surprising that the majority of these experts (approximately 70%) had a university background. Table 3 offers an overview of the experts who were consulted about test item content.

Table 3

*Distribution of experts according to fields**

University Background	**N**	**Percent**	**Outside University**	**N**	**Percent**
Teachers/professors/ researchers	299	*23.7*	High school principals/ teacher trainers	3	*0.2*
Graduate students/ doctoral candidates/post docs	35	*2.7*	Teachers	220	*17.5*
Undergraduate students	543	*43.1*	Translators	8	*0.6*
			Other academic areas	46	*3.7*
			Non-academic subject matter experts	107	*8,5*
Total	877	*69.5*		384	*30.5*

Note. *Assignment to the categories was based on information given by test developers, which sometimes was inconclusive and therefore made assignments somewhat arbitrary. Nevertheless, this table gives a general impression of the number and background of those experts.

Here, an impression is given of the cost arising from consulting experts regarding content validity. It can be assumed that it takes an expert approximately one hour to answer one question

[20] Types no. 4 d, f; 5 ab, 5b, 7h, 8c, d, e, g.
[21] The figures in Table 2 are calculated according to the frequencies given in Table 1, col. 4.
[22] The number of experts consulted ranged from one to 306 (Appendix 1, no. 67 and 17 respectively): 1-9 experts included in 21 projects, 10-49 experts in 13 projects, and more than 50 experts in 6 projects.

concerning one test instrument.[23] *From the data (Appendix 1, a combination of col. 4 and 7) the number of questions posed to experts across the 48 projects in which they were involved is known, and the number of hours of work required by them totals 4,439.*[24] *Usually experts do not receive any monetary compensation for their work. In none of the reports in our sample was a fee for the work of experts mentioned. This might be plausible with respect to university students because one might surmise optimistically that by analysing test items they achieve a certain learning gain as remuneration for their work. However, even if the amount of their contribution is subtracted, 2,177 working hours of non-student experts are left (corresponding to 272 eight-hour days of work or approximately 54 weeks of work, i.e., more than one year of full-time work). This amount of work was done almost completely by academics either at or outside the university. All of the latter were probably used to being paid relatively well. Therefore, in monetary terms their contribution was considerable.*

In pilot studies and empirical investigations, it is common to describe in detail the procedures followed to draw samples. One might expect the same applies to selecting experts for their input on test content. This expectation seems to be justified, especially since their work is essential to the quality of a newly developed measurement instrument which may be administered to thousands of test takers, and test results may be used as an important basis for decisions affecting numerous stakeholders. Therefore, I investigated whether the test developers were aware of the area(s) of expertise of their potential experts and what method they used to recruit them. The findings are listed in Appendix 1, col. 3(b, c) and are summarised in Table 4.

Table 4

*Descriptions given of procedures for selecting experts**

Content of reports on inclusion of experts	***No. of studies****
Both ➢ *expertise needed and task(s)assigned* (3b**) as well as ➢ *criteria and method(s) for selecting experts* (3c**) are reported	*13*
Only ➢ *expertise needed and task(s)assigned* (3b**) are reported	30
Only ➢ *criteria and method(s) for selecting experts* (3c**) are reported	1
➢ Neither of these is reported	12

Note. * of the 56 projects involving experts; ** see headings in Appendix 1, col. 3.

In most of the reports, test developers reported at least some idea about the expertise they needed. This was obvious because they described the tasks they gave to the experts and/or the questions they asked them. In 14 reports the test developers mentioned something vaguely about how they selected

[23] This depends on the number and type of items of the respective instrument, which can be within a broad range.
[24] In seven of these projects only the number of experts consulted was mentioned, but not the number of questions posed to them. In these cases, I surmised that only one question was posed. In nine projects the number of questions posed to experts was mentioned but not the number of experts consulted, and therefore these were omitted from this calculation.

their experts; in 13 of those, this was done after mentioning the expertise needed. In 12 projects involving experts nothing was mentioned about this issue.[25]

In addition, in empirical studies it is standard procedure to report on the results yielded from the pilot sample. Likewise, test developers could provide a qualitative report on the feedback they received from their experts and also a summative, quantitative report, especially when multiple experts were involved. In Table 5 is an overview of the reports in which experts' feedback was described (cf. Appendix 1, col. 6).

Table 5

*Reports on experts' feedback**

Kind of reports on results of expert consultation	***No. of studies****
➢ Both *qualitative* (6a**) as well as *quantitative* (6b**)	*4*
➢ Only *qualitative (6a**)*	4
➢ Only *quantitative (6b**)*	7
➢ Neither of these is reported	41

Note. * of the 56 reports including experts; ** see heading in Appendix 1, col. 6.

It should be kept in mind that test developers often are not inclined to give a complete account of their endeavours to ensure content validity with the aid of experts (Berk, 1990). Only one report in the sample (Appendix 1, no. 6) provided full details about how experts were involved according to the quality criteria explained above (cf. Table 4). Three other reports (Appendix 1, no. 17, 50, 55) come reasonably close to this standard, which is by no means excessively high, but rather customary in all other empirical issues. However, the question arises as to whether and when experts should be consulted at all when developing tests.

5. Discussion: Who is an expert and what can he or she contribute to ensure content validity?

The analysis of the outcomes of typical empirical studies of test development revealed that no recognised standard has been established on what and how to report on the involvement of experts in the development of measurement instruments. Rather, authors tend to deal unsystematically with such matters, and if they include details, they tend to be incomplete. One might assume that in some cases such details are omitted from reports due to limited space; however, a few brief details on this procedure could be expected. In our sample, such details were not provided in any report. However, this problem may also be judged differently in the editorial boards of the relevant journals.

[25] I interpreted the texts to the best of my ability.

The obligation to report on experts' contributions is one formal aspect of treating matters concerning the involvement of experts in test development. More important and particularly fundamental are the aims of involving them. From the overview given in the preceding section, it is clear that the inclusion of experts to ensure content validity of test items or tests as a whole cannot guarantee that this objective is reached.[26] The main issues are as follows: (1) It is not always clear whether the advice of an expert or even multiple experts is necessary and/or could be helpful at all. (2) "Experts" usually are not individually selected by controlling for their expertise/qualification. (3) Experts also are not assessed on the effort they make to answer the questions posed to them. (4) Variations or even contradictions in the answers of experts often are not carefully analysed, but rather are levelled out by procedures of averaging or excluding minority votes, regardless of the type of answers given (cf. Table 1). (5) Revised test items are often not returned to experts for review. (6) It remains unclear whether and to what extent the expertise of experts may be overruled by the expertise of the test developer(s). In the following, these issues are discussed in more detail.

5.1 Asking experts questions requiring truth-apt answers

In general, questions asking for truth-apt answers could be dealt with by one competent expert if not by the test developer, who usually is an expert him or herself (cf. App. 1, no. 53, 72). As mentioned above (Section 3), it is necessary to differentiate between information which is already objectively known (denoted by T_o in Table 1) and information which is still unknown but in principle could be known after having been investigated, usually by researchers (denoted by T_u in Table 1). The following must be considered:

(a) If, for example, the mathematical correctness of an attractor in an MC item is to be examined or if it is necessary to know whether certain content is part of a relevant regular curriculum, the answer can be true or false only (T_o; cf. no. 4e, 4g 5b, 7f, 7g, 8c, 8d in Table 1). In clear cases like these it does not make sense to ask multiple experts. If there is one and only one true answer, and if only highly competent experts are involved, the test developer will get only one answer from all experts consulted. Otherwise, not all experts consulted are real experts, and their dissent might be used to exclude those who provide incorrect answers from further questioning (Grant & Davis, 1997, p. 273, col. 2). Obviously, it would be a mistake to understand the dissent of experts on questions such as these as a matter of differing opinions or positions, and then to dissolve it by having a majority vote or something similar.

(b) The situation is different if, for example, the frequency or likeliness of the appearance of a particular vocational operation in a specific domain is at issue. Usually, the determination of quantities of this type (no. 4f, 6, 8g in Table 1), though truth-apt, might be unknown and will be a matter of further research (T_u). Experts normally do not initiate research projects to find answers to test developers' questions; rather, they will give an estimation of how the correct answer might read. In this case, it might be advisable (but not imperative) to ask several experts to give their *estimation* of the correct answer, and then to take the mean, mode, or another reasonable average value from their specifications. However, if the test developers have any valid information or a justified opinion of the range within which the experts' answers should be, they must exclude some answers as outliers,

[26] This question clearly differs from the problems that arise in connection with the statistically analyzable selection of items or the assessment of their reliability. Experts can also be consulted for this. However, such aspects are not the focus of this paper.

unless these alone would justify a study of this particular question to be launched. Caution should be exercised if a test item asks the test taker for his or her assumption about what others (including experts) believe to be the frequency or likeliness of this type of issue. Again, this could be known if investigated; however, this finding normally is not available. Therefore, cases like these also fall into the category T_u (see 4f in Table 1).

(c) Another type of information requested from experts requires separate consideration. Imagine, as is often the case, a question posed to experts about whether a certain distractor in an MC item will be considered plausible or implausible by test takers. With respect to one particular test taker, the expert's response (yes or no) is truth-apt but the expert usually does not know if the answer to the question actually is true or false (therefore: T_u). However, for numerous anonymous test takers an answer is even more tricky because some of the test takers may perceive that distractor as plausible and some of them may not. The answer also is unknown (T_u) as in (b) but, even worse, it will arise from a process of comparatively higher inference than in (b). The expert has to speculate on the *type of test taker* who considers the distractor to be plausible and to estimate the *percentage of test takers* of this type in the unknown multitude of possible test takers. In this case, an expert's reaction to the test developer's question is much more questionable than when speculating about only one single test taker (and all the more precarious than his or her answer to a question concerning the frequency or likeliness of observable occurrences of vocational operations as in (b)). In any case, experts' answers to questions like this will be highly uncertain and therefore not really useful in terms of ensuring content validity. Therefore, this type of question should not be posed to experts at all.

(d) Criteria of type no. 4a through d (Table 1) are expressions of an inevitable subjective valuation (S). This follows from the fact that two experts in the same domain or occupational activity may judge the relevance, importance, representativeness, or dangerousness of a given task differently, but their answers cannot be proven to be right or wrong. Even if the experts' answers to this type of question are the same, they only demonstrate consensus in their personal feelings (as viewers of a piece of artwork might do). However, unity in feelings does not change the logical character of their statements as valuations.

(e) If experts are asked to predict future developments or events (e.g., by evaluating the correctness/falsity of an answer concerning an MC item; no. 7d, Table 1), later on their answers might turn out to be true or false (as in (a) above). In this case the correct answer is not yet known, not even by researchers, and therefore cannot be given at the time the question is posed. Thus, it would be senseless to ask experts questions concerning the quality of such an answer option. MC item options of this type are flawed and should never be included. However, the situation is different if test items request forecasts based on theoretical models. For example, nobody knew in 2007 what the consequences would be if a very large bank were to go bankrupt. Nevertheless, it is imaginable that developers of tests of economic understanding might have posed a question like this to experts at that time. If experts were asked what they thought the most likely course of economic development would be, in the respective item stem there should have been the constraints *ceteris paribus* ("all else being equal") or *rebus sic stantibus* ("things thus standing"), which is common in economics and law. Further, if relying on a particular theory of international trade (e.g.,

in the tradition of the famous model of Ricardo) the correct answer could have been known (T_o).[27] For some test developers, this is reason enough to consult several or even numerous experts. Again, however, one competent expert (i.e., a person with knowledge in the respective domain) should be able to tell whether an option offered in an MC item like this is correct or incorrect. This case is either T_o or T_u depending on whether completely all (T_o) or only some (T_u) conditions/variables of the further development of circumstances are given. For a competent test developer, experts' assistance is not necessary with T_o and not helpful with T_u, because in the latter case in principle no future development can be excluded.

Initially, one could conclude that T_o items generally do not need to be judged by experts unless the test developer feels the need to verify that his or her solution is correct. In this case input from one expert should be enough. Conversely, T_u items should be excluded from all tests because their content deals with uncertain circumstances. Test developers and test users must know that, as a matter of principle, in these cases the content of the answers given by test takers (as well as by experts) cannot be judged as right or wrong.

It is quite another issue if not the knowledge of test takers, but rather their ability to handle uncertain situations, is to be assessed. Their responses to T_u items then might be judged as being acceptably precise, or more or less circumspect, or as being within an acceptable amount of time or indexed by an acceptable caveat of accuracy, and so on. This leads to another type of information to be obtained from experts: their opinion about the appropriateness of an item to stimulate test takers to show their competence in dealing with the various dimensions of uncertainty (no. 8a, 8b in Table 1).

5.2 Asking experts for subjective reactions

The alternative to asking experts for truth-apt statements is to ask them for statements indicating their approval or disapproval (acceptance or rejection) of an item, also referred to as consent statements.[28] The basis of such statements is not objectivity in the sense usually meant when referring to interpersonal verifiability. Rather, these statements are based on subjective mental states, subjective preferences, subjective appraisals or something similar. They can be identified easily by their indispensable recourse to the speaker him- or herself, even if this recourse is not made explicit. For example, if an expert says: "With respect to criterion *x* item no. *i* is fine,"[29] he or she means: "This is item no. *i* and looking at it with respect to criterion *x* my impression is positive." We also refer to such judgements as valuations.

Although valuations are subjective in character, they are different from norms (i.e., demands or claims[30]). Declarations of this type are usually followed by an exclamation mark. They consist of two elements: a description of a state of affairs as the object of an order and the order itself, such as "This is a certain state of affairs and I demand that it be maintained (established or eliminated)".[31] In the context of our literature review we may call these statements "decisions" because a decision is an explicit or

[27] See for example the Test of Economic Literacy (Third Edition) by Walstad and Rebeck (2001) including several items on further economic development subject to the conditions of the Theory of Competitive Markets.
[28] Languages comprise several types of statements (e.g., interrogative, imperative, exclamative). As differentiated in linguistics *truth-apt* statements and statements *requiring consent* are subgroups of declarative statements.
[29] Unfortunately, statements like this may be interpreted as relaying a truth-apt message such as "Item no. *i* is correct" (in the sense of no. 7f in Table 1). This is due to the polysemy of most linguistic signs. Additionally, the grammar of Indo-European languages allows a speaker to use exactly the same grammatical structure to utter a truth-apt and a consent statement (e.g., "This is green." and "This is wonderful."). Usually, the situational context eliminates this semantic deficit, but in written scientific texts it is necessary to make use of distinct and unequivocal formulations.
[30] In linguistics, referred to as *commands*, i.e., imperative sentences.
[31] E.g.: "This is item no. *j* and looking at it in respect to criterion *y* I strongly recommend it be reformulated!"

implicit statement about a personal will, wish, or at least hope.[32] By contrast, valuations do not necessarily imply a normative claim.[33]

(f) In the course of test development experts often are asked for statements indicating their approval (Grant & Davis, 1997, p. 272, col. 1; Table 2, col. 3). For example, they are asked to decide whether a particular item belongs to the domain a test is aimed at. Without an exact definition of the notion *domain* and a clear criterion for the delimitation of domains, this statement cannot be true or false (Shavelson, Gao, & Baxter, 1995). When answering such questions experts might be led by their intuitions,[34] which undoubtedly differ among individuals. In Table 1 is a list of various situations in which experts are asked or forced to make a decision.[35] The vexing problem for the test developer then is to decide how varying decisions of a group of experts should be treated. Should the test developer follow the majority? How many experts are needed to get a substantial contribution to the decision on the content validity of an item? Why should the majority decision or any decision be considered the best? Should the test developer ask the experts not only to make a decision but also to disclose the criterion on which their decision is based? What if the experts cannot name that criterion? What if the experts can, and the developer gets different decisions based on different criteria? Would it be a good idea to keep only those items on which experts are in complete agreement?

As an example, Huang and Lin (2015; cf. Appendix 1, no. 71) developed their Inventory for Measuring Student Attitudes Toward Calculus for college students. In working on content validity, they asked experts to judge the clarity of the 24 items on the test. One of the items read: "I think it is difficult to learn calculus." Is the criterion of clarity fulfilled? How would you decide? Yes or no? The experts decided "no". Therefore, the item was reworded to: "Calculus is not difficult for me." Do you agree that clarity is now given? The group of experts consisted of "three educators with expertise in instrument development, one psychological counsellor, and two professors who have each taught calculus for more than 10 years" (2015, p. 113, col. 1-2). The authors did not report on the number of experts' decisions on clarity after rewording the item. However, it is plausible that not all of them responded in the same way.

The question posed to the experts by Huang and Lin does not solicit a truth-apt answer. Judging clarity in the context of item construction obviously is a matter of subjectivity because clarity is a two-place predicate meaning "clarity for someone". The example above deals with clarity for college students, but why did Huang and Lin not ask college students, who are the genuine experts, to answer this question? Indeed, the two authors conducted a pilot study with a sample of students, but they did not ask them to judge the items on the criterion of clarity. Rather, they administered their test to them. However, to get the necessary information on clarity they should have asked the students which of the two alternative wordings to them was clearer. What would the consequences have been if they had asked students (as experts) in this way, and if they had not obtained a clear majority for one of the two alternatives? If one of the two alternatives had been favoured, would

[32] Thus, definitions are nothing other than conditional norms originating from a personal decision about the meaning of a certain term supplemented by the condition if other members of the given language community are ready to join this proposal.
[33] This becomes clear if one thinks of a person who valuates the length of days as being too short or the utility of a certain test as being unnecessary.
[34] Intuitions, in this case, may be based on and chosen from a large number of possible criteria. Even if the experts agree on those criteria, verbalised intuitions are nothing other than the expression of individual preferences.
[35] No. 1, 2, 4c, 7d.

the answers given by the minority have been considered as valid as those of the majority? Would it have been necessary to investigate relevant psychological differences between members of the minority group and those of the majority group? How could this have been done? Would validated tests have been needed to detect and identify these differences? Hence, is the result of validating test items an infinite regress? Finally, what does it mean for the validity of test interpretations if experts find a particular item only partly clear? Would it be fair to force all future test takers to respond to such an item? How many items would remain in an item pool if all those items failing to gain consensus among experts with regard to clarity (or any other predicate of the same kind) were removed?

Pragmatic issues may come into play such as the costs and benefits of developing a test in terms of money, energy, and time. How much effort has to be invested to ensure the content validity of items with respect to clarity (as in the example above) and all other "S aspects" (in accordance with Table 1 and Appendix 3)? Do financial constraints diminish the quality of test content or corrupt it totally? In other words, is content validity a quality which may vary gradually (as has been the prevailing view for a long time; Gay 1980)?[36] Can it be assumed that if this is so, the content validity correlates (strongly) with the risk of inadequate, or rather, invalid, test interpretations?

Thus, the value of experts' decisions on item validity is debatable. From what has been mentioned so far, the experts base their decisions on definitions (von Savigny, 1971), expected effects of items on test takers, or other causal relations. In the latter case, it does not make sense to ask experts for judgements, because they are overtaxed by the expectation to give an appropriate answer to a question requiring empirical knowledge (T_o). As mentioned in the example above, often the test takers are the true experts (see Table 1, no. 5a); however, in test development they usually are treated as test subjects within a pilot study, which is different from using them as experts for their decision on the quality of test items (Vogt, King, & King, 2004; Grant & Davis, 1997, p. 273, col. 2).

(g) Finally, asking experts to give *valuations* seems to be the most common way to involve experts in ensuring content validity. Experts are usually presented with scales (often Likert type) on which they state their personal valuations of a certain criterion to be adopted for a test item. For example, experts are asked to rate the relevance of an item (no. 4a in Table 1), its level of difficulty (no. 5aa), its understandability (no. 7b), its appropriateness (no. 5ad), and/or the extent to which they fully agree/agree/partly agree/disagree with a statement characterising the respective item under a certain condition (e.g., to be gender neutral (no. 5ab) or culturally unbiased (no. 5ac)).

Afterwards, the experts' answers are usually mapped numerically, presuming that they vary continuously and therefore can be averaged arithmetically. Then, the resultant value is re-translated empirically and interpreted according to the description of the grades of the given scale. If this value falls into the range of a rather positive valuation (probably higher than the arithmetic mean)

[36] In the 1980s a method called the *index of item-objective congruence* was developed for dealing with multiple experts' judgements (Thurn & Dietz 1989, referring to Rovinelli & Hambleton, 1977). This index is calculated on the basis of qualitative expert judgments (an item is "definitely" a measure of a domain, "definitely not" a measure of a domain, or no decision on this question is possible: +1, 0, -1) resulting in a measure ranging from +1.0 to -1.0. The authors suggest a minimum index of +0.7 for good items (1989, p. 342), but they do not discuss whether or to what extent index values < 1.0 diminish the quality of decisions based on items with an index value of 0.7.

the respective item, in the absence of any other obstacle, will be accepted by the test developers. Otherwise, this item will be modified or even omitted.

Although this procedure generally is considered appropriate, on closer examination it is anything but meaningful for various reasons.

aa) Often it is far from clear what the differences between full grades on a given scale mean in terms of item quality. This is all the more true for values lying between full grades, as is usually the case for the arithmetic mean. For example, imagine that (as in Jenßen, Dunekacke, and Blömeke 2015, pp. 22-23; Appendix 1, no. 6) 24 experts are asked to rate whether a given item is a good representation of all theoretically possible items. What does it mean if on a four-point scale of 1 *not at all*, 2 *rather no*, 3 *rather yes* and 4 *totally* one item gets a mean of 2.4 and a second item a mean of 2.8? In this case, the authors decided to omit the first one because its mean lies beyond the middle of the scale, that is, 2.5, whereas the second item has been kept but had to undergo extensive revision. Thus, the difference of four-tenths between 2.4 and 2.8 is significant but the same difference between 3.5 and 3.9 does not lead to any consequences. Obviously, in this case, the requirement of equidistance of data is not considered adequately.

ab) As an alternative one could consider whether the mode or the median would offer better interpretations of the experts' answers. Both have a relatively clear meaning as they mirror a verbalised subjective feeling in standardised (i.e., not a real individual's) wording. This is an advantage over the computation of an arithmetic mean usually resulting in a decimal which does not represent any of the answers given by experts. However, if the modal value exceeds the value next to it by one or two only,[37] it might be difficult to decide that this small deviation justifies the decision to prefer the mode as an adequate interpretation of their opinion. Moreover, if the median value resulting from the experts' answers lies between the next lower value on the left and the next higher value on the right, one could hardly reason that this value represents the average opinion of the experts. Choosing one of either measures would exclude all divergent interesting or relevant reasons the experts might have had in mind when answering test developers' question(s).

ac) The question remains as to whether it makes any sense at all to reduce the valuations of experts to an average. First, test developers may deal with arithmetic means quite arbitrarily. In addition, differing expert valuations may be of different quality due to judgement tendencies known from research on rating biases (e.g., leniency, central tendency, halo effect, etc.; cf. Beck, 1987, pp. 184-186), and controlling for this is very time-consuming and expensive. Furthermore, one expert might have better reasons for his or her valuation than another, an issue that refers to the fact that even valuations may include a more or less distinct portion of cognition emerging, for example, from relevant experience[38] or knowledge. Moreover, the answers experts give might be influenced by their motivation to cooperate (which might be influenced by monetary incentives), their

[37] Because the number of experts involved often is below 20 (see Appendix 1, col. 4), it is not unlikely that a result like this can be reached.

[38] It is often assumed that the longer the duration of service teachers have behind them the higher is their expertise. But, as Siedentop and Eldar (1989) have convincingly argued, this is by no means certain. Duration of experience is not a reliable indicator for expertise.

commitment to supporting the test developers, or the time they are prepared to invest in the processing time of the tasks they were given from possibly completely unknown people (i.e., the test developers) whom they are not obliged to help.

6. Conclusion

To summarise, the role of experts in ensuring the validity of the content of tests and the methods used by test developers to obtain experts' feedback is often dubious. Experts' involvement in the process of test development can vary significantly. Usually, more than one expert is consulted and often experts do not agree in their valuations, decisions, or even truth claims. Thus, the test developers have to reclaim responsibility for ensuring content validity, which they initially intended to transfer to their experts. Perhaps in an attempt to avoid giving responsibility to experts who differ in their judgements, often without being aware of it, test developers try to "average out" the procedures discussed above, which they believe permit interpretation of experts' incommensurable feedback. Moreover, if test developers are lucky enough to have consensus among their experts, they still cannot be sure the content of their test is valid because consensus is no guarantee for correctness in any sense (Popper, 1972).

Finally, test developers often do not seem to have a clear understanding of what they are asking experts to judge (e.g., the colloquial, and therefore ambiguous, terms such as relevance, difficulty, and clarity of test items). In addition, they usually do not have clear criteria against which they can assess the knowledge of external experts. Test developers might feel tempted to make use of experts' competences to legitimise their claim that their test is of good quality. However, in many cases it is the test developers themselves who are the best experts, and as such they do not really need to get advice from others on the many questions which are routinely posed to so-called experts in the procedures of test development and, in particular, of ensuring content validity. If there is a real need for advice from experts, selecting them is not easy and needs much more diligence than is routinely applied (Grant & Davis, 1997, p. 270, col. 1-2). Moreover, formulating adequate questions for experts is even more challenging.

It would be a grave misunderstanding to believe that getting experts involved in an attempt to ensure content validity is a matter of representativeness in any sense. Rather, it is a matter of careful selection of people (or a single person) having the competence to contribute to the quality of a test. Test developers should consider themselves lucky if they have tests at hand that allow them to identify competent experts. As mentioned above, this idea would lead immediately to an infinite regress.

If test developers feel it necessary to consult experts, there should be no reason not to publish the names of those experts along with the names of the test developers or at least, with their consent, to communicate their identity and lift the verbal veil of anonymity associated with the sweeping talk of "experts".[39] Thus, they not only would share responsibility but also receive appropriate recognition for their contribution to the resultant product.

The Standards for Educational and Psychological Testing (AERA, APA & NCME, 2014) highlight the importance of reporting on the process of developing tests and are guidelines with which test developers are expected to comply. It seems that editors and reviewers of journals pay little attention to the details about how experts contribute to test development, even though Standard 1.9 gives quite helpful hints on *how to report* on the inclusion of experts in the different stages of test development (not only to ensure content validity; p. 25 col. 2-26, col.1) and Standard 7.5 states "test documents should record … the nature of judgements made by subject matter experts (e.g., content validation linkages)"

[39] For example, Walstad and Rebeck (2001) give the names and affiliations of all experts involved in the development of their "Test of Economic Literacy" (pp. 3-4, 68).

(p. 126, col. 2). However, the Standards do not go into the details concerning the role experts play in providing evidence of content validity and do not elaborate on the procedures, problems, or fallacies surrounding the selection of experts and interpreting and employing their advice (pp. 14-15). With that in mind, one might conclude that further discussion is needed on the guidelines and principles to be observed when making use of the advice of experts to ensure the validity of the content of tests.

Key points

- The paper offers a comprehensive overview of experts' involvement in ensuring content validity of tests.
- A review of reports on test development (N = 72) reveals a lack of information regarding the process of selecting experts and the methodological treatment of their qualitative and quantitative input.
- A discussion on whether and when to consult or not to consult experts to ensure the content validity of tests.
- The current Standards for Educational and Psychological Testing (AERA, APA & NMCE, 2014) should be enhanced by offering detailed guidelines regarding the involvement of experts in test development to ensure content validity.

Acknowledgements

I would like to thank two anonymous reviewers and the editor who provided me with valuable advice on the presentation of my findings.

References

AERA, APA, & NCME (1985). *Standards for Educational and Psychological Testing*. Washington: APA.

AERA, APA, & NCME (2014). *Standards for Educational and Psychological Testing*. Washington: AERA.

Allen, M. J., Yen, W. M. (2002). *Introduction to measurement theory* (2nd ed.). Prospect Heights, IL: Waveland Press.

Anastasi A., Urbina S. (1997). *Psychological testing* (7th ed.). New York, NY: Prentice Hall.

Anderson, D., Irvin, S., Alonzo, J., & Tindal, G. A. (2015). Gauging Item Alignment Through Online Systems While Controlling for Rater Effects. *Educational Measurement: Issues and Practice*, 34(1), 22–33.

Angoff, W. H. (1988). Validity: an evolving concept. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 9–13). Hillsdale, NJ: Lawrence Erlbaum.

Baker, E. (2013). The Chimera of Validity. *Teachers' College Record*, 115(9).

Beck, K. (1987). Die empirischen Grundlagen der Unterrichtsforschung [The empirical foundations of research on classroom teaching. A critical analysis of the descriptive power of observation methods]. Goettingen: Hogrefe.

Beck, K., Landenberger, M. & Oser, F. (Eds.) (2016). Technoligiebasierte Kompetenzmessung in der beruflichen Bildung [Technology-based measurement in vocational education and training]. Bielefeld: Bertelsmann.

Berk, R. (1990). Importance of expert judgment in content-related validity evidence. *Western Journal of Nursing Research*, 12(5), 659–671. DOI: org/10.1177/019394599001200507

Brennan, R. L. (2013). Commentary on "Validating the Interpretations and Uses of Test Scores". *Journal of Educational Measurement*, 50(1), 74-83.

Dodd, J. (2002). Truth. *Analytic Philosophy*, 43(4), 279-291. DOI: 10.1111/1468-0149.00270 [https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-0149.00270]

Ericsson, K. A. & Smith, J. (1991). Prospects and limits of the empirical study of expertise: an introduction. In K. A. Ericsson & J. Smith (eds.), *Toward a general theory of expertise* (pp. 1-39). New York: Cambridge Univ. Press.

Field, J. (2013). Cognitive validity. In A. Geranpayeh & L. Taylor, *Examining Listening. Research and practice in assessing second language listening* (pp. 77-151). Cambridge, UK: Cambridge Univ. Press

Gay, L. R. (1980). *Educational evaluation and measurement: Competencies for analysis and application*. Columbus, OH: Charles E. Merrill.

Jackson, F., Oppy, G. & Smith, M. (1994). Minimalism and truth aptness. *Mind*, 103(411), 287-302. [https://www.jstor.org/stable/2253741?seq=1#page_scan_tab_contents]

Grant, J. S. & Davis, L. L. (1997). Selection and Use of Content Experts for Instrument Development. *Research in Nursing & Health*, 20, 269–274.

Guion, R. M. (1977). Content validity: the source of my discontent. *Applied Psychological Measurement*, 1(1), 1-10. DOI.org/10.1177/014662167700100103

Henig, J. R. (2013). The Politics of Testing When measures "Go Public". *Teachers' College Record*, 115(9), 1-11.

Kane, M. T. (2013a). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, 50(1), 1–73.

Kane, M. T. (2013b). Validation as a Pragmatic, Scientific Activity. *Journal of Educational Measurement*, 50(1), 115–122.

Kerlinger F. N. (1986). *Foundations of behavioral research* (3rd ed.). New York, NY: Holt, Rinehart, & Winston

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, 28, 563–575.

Lynn, M. (1986). Determination and quantification of content validity. *Nursing Research*, 35, 382–385.

Maas, Van der, H. L. J., Kan, K.-J. & Borsboom, D. (2014). Intelligence Is What the Intelligence Test Measures. Seriously. *Journal of Intelligence*, 2(1), 12-15. DOI: https://doi.org/10.3390/jintelligence2010012

Messick, S. (1987). *Validity*. ETS Research Report Series. Vol. 1987, Issue 2, 1-108. http://onlinelibrary.wiley.com/doi/10.1002/j.2330-8516.1987.tb00244.x/abstract; date accessed 2018/05/06; doi: 10.1002/j.2330-8516.1987.tb00244.x)

Messick, S. (1990). *Validity of Test Interpretation and Use*. ETS Research Report Series. https://eric.ed.gov/?id=ED395031; date accessed 2019/03/30.

Newton, P. E. & Shaw, S. D. (2014). *Validity in Educational and Psychological Assessment*. Los Angeles: Sage.

Pant, H. A., Zlatkin-Troitschanskaia, O., Lautenbach, C., Toepper, M. & Molerov, D. (eds.) (2016). *Modelling and Measuring Competencies in Higher Education – Validation and Methodological Innovations (KoKoHs) – Overview of the Research Projects* (KoKoHs Working Papers, 11). Berlin & Mainz: Humboldt University & Johannes Gutenberg University. http://www.kompetenzen-im-hochschulsektor.de/617_DEU_HTML.php; date accessed 2018/03/20.

Popper, K.R. (1972). *Objective Knowledge*. Oxford: Clarendon.

Reynolds C. R., Livingston R. B., Willson V. (2009). *Measurement and assessment in education* (2nd ed.). Upper Saddle River, NJ: Pearson.

Rovinelli, R. J. & Hambleton, R. K. (1977). On the use of content specialists in the assessment of criterion-referenced test item validity. *Dutch Journal for Educational Research*, 2, 49-60.

Savigny, von, E. (19712). *Grundkurs im wissenschaftlichen Definieren [Basic course on scientific defining]*. München: DTV.

Shavelson, R. J., Gao, X. & Baxter, G. P. (1995). On the content validity of performance assessments: Centrality of domain specification. In M. Birenbaum & F. Douchy (eds.), *Alternatives in Assessment of Achievements, Learning Process, and Prior Knowledge* (pp. 131–141). Boston: Kluwer Academic.

Shepard, L. A. (2013). Validity for What Purpose? *Teachers' College Record*, Vol. 115(9), p. 1-12. http://www.tcrecord.org ID Number: 17116, date accessed: 2019/01/23.

Siedentop, D. & Eldar, E. (1989). Expertise, experience, and effectiveness. *Journal of Teaching Physical Education, 8*, 254-260.

Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research*, 45(1), 83-117. DOI:org/10.1023/A:100698552

Smith, M. D. (2017). Cognitive Validity: Can Multiple-Choice Items Tap Historical Thinking Processes? *American Educational Research Journal*, 54(6), 1256-1287. DOI: 10.3102/0002831217717949

Thorn, D. W. & Deitz, J. C. (1989). Examining Content Validity Through the Use of Content Experts. *The Occupational Therapy Journal of Research*, 9, 334-346.

Vogt, D. S., King, D. W. & King, L. A. (2004). Focus Groups in Psychological Assessment: Enhancing Content Validity by Consulting Members of the Target Population. *Psychological Assessment*, 16(3), 231-243.

Walstad, W. B. & Rebeck, K. (2001). *Test of Economic Literacy. Third Edition*. New York: National Council on Economics Education.

Welner, K. G. (2013). Consequential Validity and the Transformation of Tests from Measurement Tools to Policy Tools. *Teachers' College Record*, 115(9), p. 1-6. http://www.tcrecord.org ID Number: 17115, date accessed: 06.05.2015.

White, M. C. (2018). Rater Performance Standards for Classroom Observation Instruments. *Educational Researcher*, 47(8), 492-501.

Wilson, F. R., Pan, W. & Schumsky, D. A. (2012). Recalculation of the Critical Values for Lawshe's Content Validity Ratio. *Measurement and Evaluation in Counseling and Development*, 45(3) 197–210.

Zlatkin-Troitschanskaia, O., Pant, H. A., Nagel, Th.-M., Molerov, D., Lautenbach, C. & Toepper, M. (Eds.) (2020). Portfolio of KoKoHs Assessemnts. Test Instruments for Modelling and Measuring Domain-specific and Generic Competencies of Higher Education Students and Graduates. Mainz & Berlin. https://www.wihoforschung.de/_medien/downloads/KoKoHs_Kompetenztest-Verfahren_Englisch.pdf

Appendix 1. Reports on test development with regard to including experts*

1	2	3	4	5	6	7
No.	**a) author(s) & year; page number(s) covering expert issues** *[if applicable]* **b) keyword(s)** **c) research project: acronym** *[only no. 1 through 22]*	**a) experts consulted?** *[y(es)/n(o)]* **b) aspects of expertise needed and task(s)** *[n: not reported]* **c) criteria & method(s) for selecting experts** *[n: not reported]*	**no. of experts included** *[n: not reported or n/a]*	**description of experts involved** *[n: not reported or n/a]*	**report on results of experts' consultation** **a) qualitative** **b) quantitat.** *[y(es)/n(o)]*	**type of tasks for experts (cf.Tab. 1: task list no.)** *[n: not reported]*
A Projects conducted in Germany on modeling and measuring competences/skills in higher education (2014 – 2015) [ordered alphabetically according to acronyms of the projects]						
1	a) Siebert et al. (2015, 265) b) academic text competences (teachers) c) **AkaTex** 2	a) y b) n c) n	n	researchers on writing	a) n b) n	4a
2	a) Lohse-Bossenz, Kunina-Habenicht, & Kunter (2013, 1549-1550) b) teachers' comp. in Pedago. Psych. c) **BilWiss**	a) y b) Delphi study on delimitation of domain c) n	48	psychologists (university); non-psychologists (university); teacher trainers	a) y b) n	1
3	a) Hammer et al. (2015) b) competence in teaching students with German as a 2nd language c) **DazKom**	a) n b) n/a c) n/a	n/a	n/a	a) n/a b) n/a	n/a
4	a) Eggert & Boegeholz (2009, 238) b) competence of students in decision making on the sustainable development of environment c) **ExMo**	a) y b) relevance, wording, phrasing of items c) members of a curriculum project on "Biology in Context"	research group + 10	researchers 10 biology teachers	a) n b) n	2 4a 7a, 7b, 7c
5	a) Fritsch et al. (2015, 10) b) vocational teachers' competence in CK and PCK: accounting c) **KoMeWP**	a) y b) appropriateness & relevance of items c) n	n	n	a) n b) n	4a, 4b, 4c

6	a) Jenßen, Dunekacke, & Blömeke (2015, 21-22) b) competence of kindergarten teachers in teaching maths c) **KomMa**	a) y b) y c) years of professional experience; supplementary qualifications in maths; number of publications n	24	15 researchers 9 teachers	a) y b) y	4a 6
7	a) Schwippert et al. (2014) b) competences in professional communication (economics vs. teaching) c) **KomPaed**	a) y b) description of competences: correct translation and relevance in pedagogical practice c) members of PIAAC consortia	n	PIAAC experts: intl. & nat.	a) n b) n	2 3
8	a) Trempler et al. (2015, 159) b) evidence-based practice: sub-competences (1) selection of information; (2) evaluation of studies c) **KOMPARE**	a) y b) generating a standard solution c) n	13	11 doctoral candidates and post-docs 2 professors	a) n b) y (just ICC)	n
9	a) Neumann et al. (2015, 471) b) mathematical competences of engineering students at the beginning of their academic studies c) **KoM@ING**	a) y b) n c) n	2	n	a) n b) y	4a
10	a) Schroeder, Richter, & Hoever (2008, 244) b) differentiating plausible and implausible information on a given scientific (psychological) textbase c) **KOSWO**	a) y b) n c) n	7	graduates of psychology	a) y b) y	7a 8b
11	a) Hartmann et al. (2015, 49) b) development of a test measuring preservice science teachers' scientific reasoning skills c) **Ko-WADiS**	a) y b) supervising item development with respect to relevance and representativeness c) n	8	7 subject matter experts 1 psychometrician	a) n b) n	4a, 4b

12	a) Bender et al. (2015) b) development of a competence model and measurement instrument for computer science teachers c) **KUI**	a) y b) n c) n	n	n	a) n b) n	n
13	a) Winter-Hölzl et al. (2015; 189) b) competence to write scientific texts (knowledge of genre) c) **LsScEd**	a) y b) items suiting test takers c) experience in counselling doctoral and habilitating researchers' writing processes	12	7 not specified 5 "renowned ed. researchers"	a) n b) n	7b
14	a) Tiede & Grafe (2015, 535) b) validate the model of pedagogical media competence c) **M^{3}K**	a) y b) n c) n	n	n	a) n b) n	1
15	a) Taskinen et al. (2017, 481) b) competency model for process dynamics and control (engineering) c) **MoKoMasch**	a) y b) n c) n	n		a) n b) n	n
16	a) Schladitz, Gross, & Wirtz (2015) b) educational research literacy c) **ProfilLe-P**	a) n b) n/a c) n/a	n/a	n/a	a) n/a b) n/a	n/a
17	a) Steuer et al. (2015, 211-212) b) competence in self-regulated learning c) **PRO-SRL**	a) y b) item significance and adequacy c) "experienced" university teachers and "excellent" designates students	306	144 university teachers 162 "excellent" students from 4 subjects (STEM, psychology, economics, electrical engineering)	a) n b) y	4b, 4c 5ad
18	a) Zlatkin-Troitschanskaia et al. (2015, 122) b) competence in business and economics c) **WiWiKom**	a) y b) n c) n	78	university teaching staff in business and economics	a) n b) n	3 4a 8f

B	**Projects of the German ASCOT program on measuring and modelling vocational competences in different occupations (2014 – 2015) [ordered alphabetically according to acronyms of the projects]**					
19	a) Wuttke et al. (2015, 195) b) problem solving competence of apprentices (industrial clerks) c) **DomPL-IK**	a) y b) adequacy of problems c) 5 years of experience as an employee + apprenticeship or final commercial degree	17	experts in the domain	a) n b) n	5aa
20	a) Walker, Link, & Nickolaus (2015, 229) b) problem solving competence in electronics and automation c) **KOKO EA**	a) y b) "validation" of problem scenarios c) n	n	trainers in companies teachers in vocational schools examiners	a) n b) n	2
21	a) Schmidt, Nickolaus, & Weber (2014, 556) b) structures of competence of car mechatronics c) **KOKO Kfz**	a) y b) "validation" of problem scenarios c) n	n	trainers in companies teachers in vocational schools examiners	a) n b) n	2 4f
22	a)Wittmann et al. (2014, 57-58, 61 FN 5, 62, 63) b) competences in caring for elderly people c) **TEMA**	a) y b) relevance of video-based items; "validation" of items and item difficulties c) n	n	n	a) n b) n	4a, 4b, 4c 5aa, 5ad
C	**Other research projects conducted in Germany on test development for use in the field of (higher) education (2014 – 2015) [ordered alphabetically according to first authors]**					
23	a) Esslinger (2015, 278) b) competence in reading	a) n b) n/a c) n/a	n/a	n/a	a) n/a b) n/a	n/a
24	a) Filipiak & Reis (2015) b) competences of teacher students in religion	a) n b) n/a c) n/a	n/a	n/a	a) n/a b) n/a	n/a
25	a) Kuhn (2014, 140-141, 163-170) b) PCK of teachers in business and economics: content	a) y b) content (PCK of business and economics teachers) c) n	11	university professors teacher trainers teachers	a) y b) n	2 4a, 4b, 4c 6 8b, 8f
26	a) Lindl & Kloiber (2015) b) PCK of Latin teachers	a) n b) n/a c) n/a	n/a	n/a	a) n/a b) n/a	n/a

27	a) Vogler, Messmer, & Allemann (2017, 336) b) PCK of sports teachers	a) y b) 3 Delphi studies: one of them on item content c) n	5	1 from abroad 1 with practical experience 1 without experience 1 with psych. knowledge 1 researcher	a) n b) n	4a
D Journal *Educational and Psychological Measurement*, section. "Validity Studies", 2009(1) – 2015(6) [ordered by release date; upwards]						
28	a) Ordoñez et al. (2009, 291) b) epistemological beliefs	a) y b) translation/back translation English – Spanish c) professional interpreters	2	translators familiar with both cultures (Anglo-Saxon & Columbian)	a) n b) n	3
29	a) Liu et al. (2009, 479) b) standardised letters of recommendation	a) y b) overall coverage and clarity of items c) n	n	n	a) n b) n	4a 7a
30	a) Lei (2009, 828) b) preschoolers' numeracy skills	a) n b) n/a c) n/a	n/a	[authors' professional experience]	a) n/a b) n/a	n/a
31	a) Aydın & Uzuntiryaki (2009, 871) b) high school chemistry self-efficacy scale	a) y b) grammar, clarity of items; content representativeness c) n	13	1 specialist in Turkish; 12 chemistry high school teachers, researchers in chemistry, chemistry educators, ed. psychologists, measurement specialists	a) y b) n	4a 7a, 7c
32	a) Hulpia, Devos, & Rosseel (2009, 1018) b) High School Chemistry Self-Efficacy Scale	a) y b) item complexity; feasibility of questionnaire c) n	n	n	a) n b) n	5aa
33	a) Cadiz, Sawyer, & Griffith (2009, 1042, 1043) b) absorptive capacity and experienced community of practice	a) y b) appropriateness of items c) n	n 185	current employees MBA students	a) n b) n	4a, 4b, 4c, 4e 7a, 7b
34	a) Fletcher & Nusbaum (2010, 108) b) competitive work environment scale	a) y b) item content, clarity, representativeness c) n	3	doctoral candidates in industrial & organizational psychology	a) n b) n	4a 7a

35	a) Luttrell et al. (2010, 146-148) b) mathematics value inventory	a) y b) clarity, technical adequacy, content c) n	5 38	experts in mathematics education students in a graduate-level measurement class	a) y b) y	4b, 4c 6 7a, 7b, 7d, 7f, 7g
36	a) Myers (2010, 482) b) athletes' perceptions of coaching competency	a) y b) adapting items from American English to British English c) n	n	n	a) n b) n	3
37	a) Mesmer-Magnus et al. (2010, 514-515) b) co-workers' informal work accommodation to family	a) y b) generate & judge 75 instances, 6 categories 31 items c) n	57	adult employees and 2 groups of subject matter experts (=graduate students) plus authors	a) n b) n	2
38	a) Linnenbrink-Garcia et al. (2010) b) situational interest in academic domains	a) n b) n/a c) n/a	n/a	n/a	a) n/a b) n/a	n/a
39	a) Teo (2010, 995) b) technology acceptance measure for preservice teachers	a) y b) clarity, conciseness c) n	20	preservice teachers	a) n b) n	7a, 7e
40	a) McDermott et al. (2011, 151-152) b) teachers' graded responses for preschoolers' stylistic learning behaviour	a) y b) finding the target structure for CFA c) n	n	n	a) y b) n	4g
41	a) Cheng et al. (2011, 202-203) b) school bullying scales	a) y b) item review c) n	7	3 professors in counselling 2 professors in education 1 professor in testing/measurem. 1 high school principal	a) n b) n	n
42	a) Gable et al. (2011, 220) b) knowledge of Internet risk & Internet behaviour	a) y b) item review c) familiar with bullying behaviour	7	5 middle school teachers 2 principals	a) n b) n	n
43	a) Nilsson et al. (2011, 261) b) social issues advocacy scale	a) y b) item review c) n	8	5 graduate students 3 faculty members	a) n b) n	4c 7a, 7b, 7c, 7d
44	a) Curşeu & Schruijer (2012, 1055) b) general decision-making style	a) y b) (back-)translation Engl.-Dutch c) bilingual people	n	n	a) n b) n	3

45	a) Warner, Koufteros, & Verghese (2014; 1000, 1006) b) second language learning aptitude in technology acceptance	a) y b) 1. computer anxiety (CA): enhance content and face validity 2. computer user learning aptitude (CULA): appropriate-ness, comprehensiveness c) n	9	academics from different disciplines experts from information systems and research methodology	a) n b) n	4a, 4b, 4c 7a, 7d 8b
46	a) Paulhus & Dubois (2014, 980) b) overclaiming technique to scholastic assessment	a) n b) n/a c) n/a	n/a	n/a	a) n/a b) n/a	n/a
47	a) Kersting, Sherin, & Stigler (2014, 956) b) rating of teacher statements; interrater reliability	a) y b) n c) n	n		a) n b) n	n
48	a) Nezhnov et al. (2015), 242-244) b) mathematical knowledge	a) n b) n/a c) n/a	n/a	n	a) n /a b) n/a	n/a
49	a) Dimitrov, Raykov, & AL-Qataee (2015, 475-490) b) general academic ability	a) n b) n/a c) n/a	n/a	n	a) n/a b) n/a	n/a
E	**Journal *Measurement and Evaluation in Counselling and Development*, section "Assessment, Development, and Validation", 2009(1) – 2015(4) [ordered by publication date; upwards]**					
50	a) Kim, Soliz Orellana, & Alamilla (2009, 75, col. 2) b) Latino/a value scale	a) y b) do items represent Latino culture; are they culturally unbiased c) bilingual people	9	3 Latina doctoral candidates 5 Latina undergraduate students 1 Asian American professor in counselling psychology	a) y b) n	2 5ac
51	a) Tovar, Simon, & Lee (2009) b) college mattering	a) n b) n/a c) n/a	n/a	n/a	a) n/a b) n/a	n/a
52	a) Pistole & Roberts (2011, 67) b) long distance romantic relationship (LDR)	a) y b) items stimulating resonance with LDR; consistency with LDR; importance to discrimina-te from GCR [Geogr. Close R] c) n	4	1 faculty member 3 graduate students, all familiar with LDR	a) n b) n	2 4a, 4b

53	a) Kopp et al. (2011, 112-113) b) academic entitlement	a) n b) n/a c) n/a	n/a	n/a [judgment about content validity done by the authors]	a) n/a b) n/a	n/a
54	a) Kim et al. (2011, 137, col. 1) b) experience with close relationships revised scale (ECRR)	a) y b) translation English to Korean c) bilingual people	n	n	a) n b) n	3
55	a) Adelson & McCoach (2011, 230-231) b) attitudes of students in grades 3 to 6 toward maths; 3 constructs: maths self-perception/enjoyment of maths/perceived usefulness of maths	a) y b) *assigning* items to constructs *rating* definitions of constructs/ wording & clarity of items c) n	17	former and actual teachers, thereof: 6 professors in mathematics education; 4 doctoral candidates 3 upper elementary students	a) y b) y	6 7a, 7c
56	a) Neto (2012) b) satisfaction with sex life scale	a) n b) n/a c) n/a	n/a	n/a	a) n/a b) n/a	n/a
57	a) Sun, Ng, & Wang (2012, 137-138) b) dispositional hope scale	a) y b) translation English – Chinese (forth & back) discussion until agreement c) bilingual people	4	bilingual English-Chinese Americans (associates of first author)	a) y b) n	3
58	a) Locke et al. (2012, 154, 155, 157, 165) b) counseling centre assessment of psychological symptoms	a) y b) judgment of psychometric properties of items and subscales c) n	12	10 senior staff clinicians of a university counselling center 2 experienced psychotherapy researchers	a) y b) n	2 4a 6
59	a) Erford & Alsamadi.C. (2012) b) emotional problems–parent report (STEP-P)	a) n b) n/a c) n/a	n/a	n/a	a) n/a b) n/a	n/a
60	a) Hardesty & Richardsin (2012) b) social support for adolescents	a) n b) n/a c) n/a	n/a	n/a	a) n/a b) n/a	n/a
61	a) Lim & Chapman (2013, 28-29) b) Fennema-Sherman Mathematics Anxiety Subscale	a) y b) suitability of items c) n	8	7 mathematics teachers 1 person from University of Cambridge Local Examinations Syndicate	a) n b) n	2 7a

62	a) Ng et al. (2013, 91, col 2) b) campus caring	a) y b) verify content and wording of items c) n	146	university students	a) n b) n	2 4a 7a, 7d
63	a) Jacobs & Struyf (2013, 161) b) socioemotional guidance at school	a) y b) content validity; difficulty, clarity, feasibility of items; translation Dutch – English c) domain and interpreter experts	142	26 teachers, guidance teachers, principals, trainers 115 teachers 1 expert in Dutch, English, and educational vocabulary	a) n b) n	5aa 7a 8a
64	a) Rantanen & Soini (2013, 253) b) response observation system measurement and evaluation in counseling and development	a) y b) validity, adequacy of categories and items c) n	3	2 doctoral students, 1 doctoral researcher	a) n b) y	4a, 4g 7e
65	a) Balkin et al. (2014; 7) b) forgiveness reconciliation inventory	a) y b) n c) n	2	experts in publications in counseling and forgiveness	a) n b) n	n
66	a) Boudreaux (2014, 16, col. 2) b) attitudes toward anger management scale	a) y b) n c) n	4	graduate students	a) n b) n	n
67	a) Montes et al. (2014, 44) b) mindful attention awareness scale	a) y b) n c) n	1	n	a) n b) n	n
68	a) Carey et al. (2014, 173) b) academic success skills	a) y b) n c) n	4	2 elementary teachers 2 school counselors	a) n b) n	n
69	a) Dominguez Espinosa & van de Vijver (2014, 202, col. 1) b) social desirability scale	a) y b) overlap with another instrument c) n	2	doctoral students	a) n b) n	2
70	a) Ahn, Ebesutani, & Kamphaus (2014, 229, col, 2) b) behaviour assessment system for children	a) y b) forth & back translation English-Korean c) interpreters	2	n	a) n b) n	3
71	a) Huang & Lin (2015, 113) b) attitudes toward calculus	a) y b) clarity, readability, appropriateness c) n	3	3 educators, psych. 1 counselor 2 professors experienced in teaching calculus	a) y b) n	7a, 7b, 8b

72	a) Jenkins-Guarnieri, Vaughan, & Wright (2015, 270, col. 2) b) self-determination measure for college students	a) n b) n/a c) n/a	n/a	[2nd author functioning also as "expert"]	a) n/a b) n/a	n/a

* chosen from two programs funded by the German Ministry of Education and Research devoted to the development of instruments for measuring competences (academic students: *KoKoHs*: Pant et al. 2016; apprentices: *ASCOT*: Beck et al. 2016) plus individual projects conducted in Germany on test development (time span 2014-2015) and two journals with a focus on the development of measurement instruments in psychology/education (time span 2009-2015): *Educational and Psychological Measurement* (Sage, Los Angeles et al.) and *Measurement and Evaluation in Counselling and Development*, (Sage, Los Angeles et al.)

° not specified in further detail

Appendix 2. Sources

A Projects conducted as part of the KoKoHs program on modeling and measuring competences/skills in higher education (2014 – 2015) [in alphabetical order according to acronym]

(1) Siebert-Ott, G., Decker, L., Kaplan, I., & Macha, K. (2015). Akademische Textkompetenzen bei Studienanfängern und fortgeschrittenen Studierenden des Lehramtes (AkaTex) – Kompetenzmodellierung und erste Ergebnisse der Kompetenzerfassung. In U. Riegel, S. Schubert, G. Siebert-Ott & K. Macha (Hrsg.), *Kompetenzmodellierung und Kompetenzmessung in den Fachdidaktiken* (S. 257-273). Münster: Waxmann. **AkaTex**

(2) Lohse-Bossenz, H., Kunina-Habenicht, O., & Kunter, M. (2013). The role of educational psychology in teacher education: expert opinions on what teachers should know about learning, development, and assessment. *European Journal of Psychology of Education*, 28, 1543-1565. DOI: 10.1007/s10212-013-0181-6. **BilWiss**

(3) Hammer, S., Carlson, S.A., Ehmke, T., Koch-Priewe, B., Koeker, A., Ohm, U., Rosenbrock, S., & Schulze, N. (2015). Kompetenz von Lehramtsstudierenden in Deutsch als Zweitsprache. In S. Blömeke & O. Zlatkin-Troitschanskaia, *Kompetenzen von Studierenden* (S. 33-54). Zeitschrift für Pädagogik. Beiheft 61. Weinheim: Beltz Juventa. **DazKom**

(4) Eggert, S. & Boegeholz, S. (2009). Students' use of decision-making strategies with regard to socioscientific issues: An application of the Rasch partial credit model. *Science Education*, 94(2), 230-258. **ExMo**

(5) Fritsch, S., Berger, S., Seifried, J., Bouley, F., Wuttke, E., Schnick-Vollmer, K., & Schmitz, B. (2015). The impact of university teacher training on prospective teachers' CK and PCK – a comparison between Austria and Germany. *Empirical Research in Vocational Education and Training*, 7(4), 1-20. http://www.ervet-journal.com/content/7/1/4; DOI: 10.1186/s40461-015-0014-8. **KoMeWp**

(6) Jenßen, L., Dunekacke, S., & Blömeke, S. (2015). Qualitätssicherung in der Kompetenzforschung. In S. Blömeke & O. Zlatkin-Troitschanskaia, *Kompetenzen von Studierenden* (S. 11-31). Zeitschrift für Pädagogik. Beiheft 61. Weinheim: Beltz Juventa. **KomMa**

(7) Schwippert, K., Braun, E., Prinz, D., Schaeper, H., Fickermann, D., Pfeiffer, J., & Brachem, J.-C. (2014). KomPaed - Tätigkeitsbezogene Kompetenzen in pädagogischen Handlungsfeldern. *Die Deutsche Schule*, 106(1), 72-84. **KomPaed**

(8) Trempler, K., Hetmanek, A., Wecker, C., Kiesewetter, J., Wermelt, M., Fischer, F., Fischer, M., & Graesel, C. (2015). Nutzung von Evidenz im Bildungsbereich. Validierung eines Instruments zur Erfassung von Kompetenzen der Informationsauswahl und Bewertung von Studien. In S. Blömeke & O. Zlatkin-Troitschanskaia, *Kompetenzen von Studierenden* (S. 144-166). Zeitschrift für Paedagogik. Beiheft 61. Weinheim: Beltz Juventa. **KOMPARE**

(9) Neumann, I., Rösken-Winter, B., Lehmann, M., Duchhardt, C., Heinze, A., & Nickolaus, R. (2015). Measuring Mathematical Competences of Engineering Students at the Beginning of Their Studies. *Peabody Journal of Education*, 90(4), 465-476. DOI: 10.1080/0161956X.2015.1068054. **KoM@Ing**

(10) Schroeder, S., Richter, T. & Hoever, I. (2008). Getting a picture that is both accurate and stable: Situation models and epistemic validation. *Journal of Memory and Language*, 59, 237-255. DOI: 10.1016/j.jml.2008.05.001. **KOSWO**

(11) Hartmann, S., Upmeier zu Belzen, A., Krüger, D., & Pant, H. A. (2015). Scientific Reasoning in Higher Education. Constructing and Evaluating the Criterion-Related Validity of an Assessment of Preservice Science Teachers' Competencies. *Zeitschrift für Psychologie*, 223(1), 47-53. DOII: 10.1027/2151-2604/a000199. **Ko-WADiS**

(12) Bender, E., Hubwieser, P., Schaper, N., Margaritis, M., Berges, M., Ohrndorf, L., Magenheim, J., & Schubert, S. (2015). Towards a Competency Model for Teaching Computer Science. *Peabody Journal of Education*, 90(4), 519-532. DOI: 10.1080/0161956X.2015.1068082. **KUI**

(13) Winter-Hölzl, A., Waeschle, K., Wittwer, J., Watermann, R., & Nueckles, M. (2015). Entwicklung und Validierung eines Tests zur Erfassung des Genrewissens Studierender und Promovierender der Bildungswissenschaften. In S. Blömeke & O. Zlatkin-Troitschanskaia, *Kompetenzen von Studierenden* (S. 185-202). Zeitschrift für Pädagogik. Beiheft 61. Weinheim: Beltz Juventa. **LsScED**

(14) Tiede, J., Grafe, S., & Hobbs, R. (2015). Pedagogical Media Competencies of Preservice Teachers in Germany and the United States: A Comparative Analysis of Theory and Practice, *Peabody Journal of Education*, 90(4), 533-545. DOI: 10.1080/0161956X.2015.1068083. **M3K**

(15) Taskinen, P. H., Steimel, J., Gräfe, L., Engell, S., & Frey, A. (2015). A Competency Model for Process Dynamics and Control and Its Use for Test Construction at University Level, *Peabody Journal of Education*, 90(4), 477-490. DOI: 10.1080/0161956X.2015.1068074. **MoKoMasch**
(16) Schladitz S., Gross, J., & Wirtz, M. (2015). Konstruktvalidierung eines Tests zur Messung bildungswissenschaftlicher Forschungskompetenz. In S. Blömeke & O. Zlatkin-Troitschanskaia, *Kompetenzen von Studierenden* (S. 167-184). Zeitschrift für Pädagogik. Beiheft 61. Weinheim: Beltz Juventa. **ProfiLe-P**
(17) Steuer, G., Engelschalk, T., Joestl, G., Roth, A., Wimmer, B., Schmitz, B., Schober, B., Spiel, C., Ziegler, A., & Dresel, M. (2015). Kompetenzen zum selbstregulierten Studium. In S. Blömeke & O. Zlatkin-Troitschanskaia, *Kompetenzen von Studierenden* (S. 203-225). Zeitschrift für Pädagogik. Beiheft 61. Weinheim: Beltz Juventa. **PRO-SRL**
(18) Beck, K., Landenberger, M. & Oser, F. (Hrsg.) (2016). Technologiebasierte Kompetenzmessung in der beruflichen Bildung. Bielefeld: Bertelsmann.. **WiWiKom**

B Projects of the German ASCOT program on measuring and modelling vocational competences in different occupations (2014 – 2015) [in alphabetical order according to acronym]

(19) Wuttke, E., Seifried, J., Brandt, S., Rausch, A., Sembill, D., Martens, T., & Wolf, K. (2015). Modellierung und Messung domänenspezifischer Problemlösekompetenz bei angehenden Industriekaufleuten – Entwicklung eines Testinstruments und erste Befunde zu kognitiven Kompetenzfacetten. *Zeitschrift für Berufs- und Wirtschaftspaedagogik*, 111(2), 189-207. **DomPL-IK**
(20) Walker, F., Link, N., & Nickolaus, R. (2015). Berufsfachliche Kompetenzstrukturen bei Elektronikern für Automatisierungstechnik am Ende der Berufsausbildung. *Zeitschrift für Berufs- und Wirtschaftspaedagogik*, 111(2), 222-241. **KOKO-EA**
(21) Schmidt, T., Nickolaus, R., & Weber, W. (2014). Modellierung und Entwicklung des fachsystematischen und handlungsbezogenen Fachwissens von Kfz-Mechatronikern. *Zeitschrift für Berufs- und Wirtschafts-paedagogik*, 110, 549-574. **KOKO-Kfz**
(22) Wittmann, E., Weyland, U., Nauerth, A., Döring, O., Rechenbach, S., Simon, J., & Worofka, I. (2014). Kompetenzerfassung in der Pflege älterer Menschen – Theoretische und domänenspezifische Anforderungen der Aufgabenmodellierung. In J. Seifried, U. Fasshauer & S. Seeber (Hrsg.), *Jahrbuch der berufs- und wirtschaftspädagogischen Forschung 2014* (S. 53-66). Opladen: Barbara Budrich. **TEMA**

C Other research projects conducted in Germany on test development in the field of (higher) education (2014 – 2015) [in alphabetical order according to first author]

(23) Esslinger, G. (2015). Kommas beim Lesen verarbeiten können – eine vernachlässigte Teilkompetenz allgemeiner Lesefähigkeit. In U. Riegel, S. Schubert, G. Siebert-Ott, & K. Macha (Hrsg.), *Kompetenzmodellierung und Kompetenzmessung in den Fachdidaktiken* (S. 274-292). Münster: Waxmann.
(24) Filipiak, A. & Reis, O. (2015). Was lernen Studierende in der Systematischen Theologie? Kompetenzdiagnostik in der Religionslehrerbildung. In U. Riegel, S. Schubert, G. Siebert-Ott, & K. Macha (Hrsg.), *Kompetenzmodellierung und Kompetenzmessung in den Fachdidaktiken* (S. 227-241). Münster: Waxmann.
(25) Kuhn, C. (2014). *Fachdidaktisches Wissen von Lehrkräften im kaufmännisch-verwaltenden Bereich. Modellbasierte Testentwicklung und Validierung*. Landau: VEP
(26) Lindl, A. & Kloiber, H. (2015). Erste Schritte zur Kompetenzmessung von Lateinlehrkräften. In U. Riegel, S. Schubert, G. Siebert-Ott, & K. Macha (Hrsg.), *Kompetenzmodellierung und Kompetenzmessung in den Fachdidaktiken* (S. 293-305). Münster: Waxmann.
(27) Vogler, J., Messmer, R., & Allemann, D. (2017). Das fachdidaktische Wissen und Können von Sportlehrpersonen (PCK-Sport). *German Journal of Exercise and Sport Research*, 47(4), 335-347.

D Journal *Educational and Psychological Measurement*, sec. "Validity Studies", 2009(1) – 2015(6) [in chronological order of release date]

(28) Ordoñez, X. G., Ponsoda, V., Abad, F. J., & Romero, S. J. (2009). Measurement of Epistemological Beliefs: Psychometric Properties of the EQEBI Test Scores. *Educational and Psychological Measurement*, 69(2), 287-302. DOI: 10.1177/0013164408323226

(29) Liu, O. L., Minsky, J., Ling, G., & Kyllonen, P. (2009). Using the Standardized Letters of Recommendation in Selection: Results From a Multidimensional Rasch Model. *Educational and Psychological Measurement*, 69(3), 475-492. DOI: 10.1177/0013164408322031

(30) Lei, P.-W., Wu, Q., DiPerna, J., & Morgan, P. L. (2009). Developing Short Forms of the EARLI Numeracy Measures. Comparison of Item Selection Methods. *Educational and Psychological Measurement*, 69(5), 825-842. DOI: 10.1177/0013164409332215

(31) Aydın, Y. Ç. & Uzuntiryaki, E. (2009). Development and Psychometric Evaluation of the High School Chemistry Self-Efficacy Scale. *Educational and Psychological Measurement*, 69(5), 868-880. DOI: 10.1177/0013164409332213

(32) Hulpia, H., Devos, G., & Rosseel, Y. (2009). Development and Validation of Scores on the Distributed Leadership Inventory. *Educational and Psychological Measurement*, 69(6), 1013-1034. DOI: 10.1177/0013164409344490

(33) Cadiz, D., Sawyer, J. E., & Griffith, T. L. (2009). Developing and Validating Field Measurement Scales for Absorptive Capacity and Experienced Community of Practice. *Educational and Psychological Measurement*, 69(6), 1035-1058. DOI: 10.1177/0013164409344494

(34) Fletcher, Th. D. & Nusbaum, D. N. (2010). Development of the Competitive Work Environment Scale: A Multidimensional Climate Construct. *Educational and Psychological Measurement*, 70(1), 105-124. DOI: 10.1177/0013164409344492

(35) Luttrell, V. R., Callen, B. W., Allen, C. S., Wood, M. D., Deeds, D. G., & Richard, D. C. S. (2010). The Mathematics Value Inventory for General Education Students: Development and Initial Validation. *Educational and Psychological Measurement*, 70(1), 142-160. DOI: 10.1177/0013164409344526

(36) Myers, N. D., Chase, M. A., Beauchamp, M. R., & Jackson, B. (2010). Athletes' Perceptions of Coaching Competency Scale II-High School Teams. *Educational and Psychological Measurement*, 70(3), 477-494. DOI: 10.1177/0013164409344520

(37) Mesmer-Magnus, J., Murase, T., DeChurch, L. A., & Jiménez, M. (2010). Coworker Informal Work Accommodations to Family: Scale Development and Validation. *Educational and Psychological Measurement*, 70(3), 511-531. DOI: 10.1177/0013164409355687

(38) Linnenbrink-Garcia, L., Durik, A. M., Conley, A. M., Barron, K. E., Tauer, J. M., Karabenick, S. A., & Harackiewicz, J. M. (2010). Measuring Situational Interest in Academic Domains. *Educational and Psychological Measurement*, 70(4), 647-671. DOI: 10.1177/0013164409355699

(39) Teo, T. (2010). The Development, Validation, and Analysis of Measurement Invariance of the Technology Acceptance Measure for Preservice Teachers (TAMPST). *Educational and Psychological Measurement*, 70(6), 990-1006. DOI: 10.1177/0013164410378087

(40) McDermott, P. A., Fantuzzo, J. W., Warley, H. P., Waterman, C., Angelo, L. E., Gadsden, V. L., & Sekino, Y. (2001). Multidimensionality of Teachers' Graded Responses for Preschoolers' Stylistic Learning Behavior: The Learning-to-Learn Scales. *Educational and Psychological Measurement*, 71(1), 144-169. DOI: 10.1177/0013164410387351

(41) Cheng, Y.-Y., Chen, L.-M. Liu, K.-S., & Chen, Y.-L. (2011). Development and Psychometric Evaluation of the School Bullying Scales: A Rasch Measurement Approach. *Educational and Psychological Measurement*, 71(1), 200-216. DOI: 10.1177/0013164410387387

(42) Gable, R. K., Ludlow, L. H., McCoach, D. B., & Kite, S. L. (2011). Development and Validation of the Survey of Knowledge of Internet Risk and Internet Behavior. *Educational and Psychological Measurement*, 71(1), 217-230. DOI: 10.1177/0013164410387389

(43) Nilsson, J. E., Marszalek, J. M., Linnemeyer, R. M., Bahner, A. D., & Misialek, L. H. (2011). Development and Assessment of the Social Issues Advocacy Scale. *Educational and Psychological Measurement*, 71(1), 258-275. DOI: 10.1177/0013164410391581

(44) Curşeu, P. L. & Schruijer, S. G. (2012). Decision Styles and Rationality: An Analysis of the Predictive Validity of the General Decision-Making Style Inventory. *Educational and Psychological Measurement*, 72(6), 1053-1062. DOI: 10.1177/0013164412448066

(45) Warner, J. A., Koufteros, X., & Verghese, A. (2014). Learning Computerese: The Role of Second Language Learning Aptitude in Technology Acceptance. *Educational and Psychological Measurement*, 74(6), 991-1017. DOI: 10.1177/0013164414520629
(46) Paulhus, D. L. & Dubois, P. J. (2014). Application of the Overclaiming Technique to Scholastic Assessment. *Educational and Psychological Measurement*, 74(6), 975-990. DOI: 10.1177/0013164414536184
(47) Kersting, N. B., Sherin, B. L., & Stigler, J. W. (2014). Automated Scoring of Teachers' Open-Ended Responses to Video Prompts: Bringing the Classroom-Video-Analysis Assessment to Scale. *Educational and Psychological Measurement*, 74(6), 950-974. DOI: 10.1177/0013164414521634
(48) Nezhnov, P., Kardanova, E., Vasilyeva, M., & Ludlow, L. (2015). Operationalizing Levels of Academic Mastery Based on Vygotsky's Theory: The Study of Mathematical Knowledge. *Educational and Psychological Measurement*, 75(2), 235-259. DOI: 10.1177/0013164414534068
(49) Dimitrov, D. M., Raykov, T., & AL-Qataee, A. A. (2015). Developing a Measure of General Academic Ability. *Educational and Psychological Measurement*, 75(3), 475-490.

E Journal *Measurement and Evaluation in Counselling and Development*, sec. "Assessment, Development, and Validation", 2009(1) – 2015(4) [in chronological order by release date]

(50) Kim, B. S., Soliz, A., Orellana, B., & Alamilla, S. G. (2009). Latino/a Values Scale. Development, Reliability, and Validity. *Measurement and Evaluation in Counseling and Development*, 42(2), 71-91. DOI: 10.1177/0748175609336861
(51) Tovar, E., Simon, M. A., & Lee, H. B. (2009). Development and Validation of the College Mattering Inventory With Diverse Urban College Students. *Measurement and Evaluation in Counseling and Development*, 42(3), 154-178. DOI: 10.1177/0748175609344091
(52) Pistole, M. C. & Roberts, A. (2011). Measuring Long-Distance Romantic Relationships: A Validity Study. Measure. *Measurement and Evaluation in Counseling and Development*, 44(2), 63-76. DOI: 10.1177/0748175611400288
(53) Kopp, J. P., Zinn, T. E., Finney, S. J., & Jurich, D. P. (2011). The Development and Evaluation of the Academic Entitlement Questionnaire. *Measurement and Evaluation in Counseling and Development*, 44(2), 105-129. DOI: 10.1177/0748175611400292
(54) Kim, S.-H., Sherry, A. R., Lee, Y.-S., & Kim, C.-D. (2011). Psychometric Properties of a Translated Korean Adult Attachment Measure. *Measurement and Evaluation in Counseling and Development*, 44(3), 135-150. DOI: 10.1177/0748175611409842
(55) Adelson, J. L. & McCoach, D. B. (2011). Development and Psychometric Properties of the Math and Me Survey: Measuring Third Through Sixth Graders' Attitudes Toward Mathematics. *Measurement and Evaluation in Counseling and Development*, 44(4), 225-247. DOI: 10.1177/0748175611418522
(56) Neto, F. (2012). The Satisfaction With Sex Life Scale. *Measurement and Evaluation in Counseling and Development*, 45(1), 18-31. DOI: 10.1177/0748175611422898
(57) Sun, Q., Ng, K.-M., & Wang, C. (2012). A Validation Study on a New Chinese Version of the Dispositional Hope Scale. *Measurement and Evaluation in Counseling and Development*, 45(2), 133-148. DOI: 10.1177/0748175611429011
(58) Locke, B. D., McAleavey, A. A., Zhao, Y., Lei, P.-W., Hayes, A., Castonguay, L. G., Li, H., Tate, R., & Lin, Y.-C. (2012). Development and Initial Validation of the Counseling Center Assessment of Psychological Symptoms–34. *Measurement and Evaluation in Counseling and Development*, 45(3), 151-169. DOI: 10.1177/0748175611432642
(59) Erford, B. T. & Alsamadi, S. C. (2012). The Screening Test for Emotional Problems–Parent Report (STEP-P). Studies of Reliability and Validity. *Measurement and Evaluation in Counseling and Development*, 45(3), 170-180. DOI: 10.1177/0748175611432643
(60) Hardesty, P. H. & Richardsin, G. B. (2012). The Structure and Validity of the Multidimensional Social Support Questionnaire. *Measurement and Evaluation in Counseling and Development*, 45(3), 181-196. DOI: 10.1177/0748175612441214
(61) Lim, S. Y. & Chapman, E. (2013). An Investigation of the Fennema-Sherman Mathematics Anxiety Subscale. *Measurement and Evaluation in Counseling and Development*, 46(1), 26–37. DOI: 10.1177/0748175612459198
(62) Ng, P., Su, X. S., Chan, V., Leung, H., Cheung, W., & Tsun, A. (2013). The Reliability and Validity of a Campus Caring Instrument Developed for Undergraduate Students in Hong Kong. *Measurement and Evaluation in Counseling and Development*, 46(2), 88–100. DOI: 10.1177/0748175612467463

(63) Jacobs, K. & Struyf, E. (2013). Measuring Integrated Socioemotional Guidance at School: Factor Structure and Reliability of the Socioemotional Guidance Questionnaire (SEG-Q). *Measurement and Evaluation in Counseling and Development*, 46(3), 159–177. DOI: 10.1177/0748175613481978
(64) Rantanen, A. P. & Soini, H. S. (2013). Development of the Response Observation System. *Measurement and Evaluation in Counseling and Development*, 46(4), 247–260. DOI: 10.1177/0748175613484041
(65) Balkin, R. S., Harris, N. A., Freeman, S. J., & Huntington, S. (2014). The Forgiveness Reconciliation Inventory: An Instrument to Process Through Issues of Forgiveness and Conflict. *Measurement and Evaluation in Counseling and Development*, 47(1), 3–13. DOI: 10.1177/0748175613497037
(66) Boudreaux, D. J., Dahlen, E. R., Madson, M. B., & Yowell, E. B. (2014). Attitudes Toward Anger Management Scale: Development and Initial Validation. *Measurement and Evaluation in Counseling and Development*, 47(1), 14–26. DOI: 10.1177/0748175613497039
(67) Montes, S. A., Ledesma, R. D., Garcia, N. M., & Poó, F. M. (2014). The Mindful Attention Awareness Scale (MAAS) in an Argentine Population. *Measurement and Evaluation in Counseling and Development*, 47(1), 43–51. DOI: 10.1177/0748175613513806
(68) Carey, J., Brigman, G., Webb, L., Villares, E., & Harrington, K. (2014). Development of an Instrument to Measure Student Use of Academic Success Skills: An Exploratory Factor Analysis. Measurement and Evaluation inCounseling and Development, 47(3), 171–180. DOI: 10.1177/0748175613505622
(69) Dominguez Espinosa, A. & van de Vijver, F. J. R. (2014). An Indigenous Social Desirability Scale. *Measurement and Evaluation in Counseling and Development*, 47(3), 199–214. DOI: 10.1177/0748175614522267
(70) Ahn, C. M., Ebesutani, C., & KamphausR. W. (2014). A Psychometric Analysis and Standardization of the Behavior Assessment System for Children-2, Self-Report of Personality, College Version, Among a Korean Sample. *Measurement and Evaluation in Counseling and Development*, 47(3), 226–244. DOI: 10.1177/0748175614531797
(71) Huang, Y.-C. & Lin, S.-H. (2015). Development and Validation of an Inventory for Measuring Student Attitudes Toward Calculus. *Measurement and Evaluation in Counseling and Development*, 48(2), 109–123. DOI: 10.1177/0748175614563314
(72) Jenkins-Guarnieri, M. A., Vaughan, A. L., & Wright, S. L. (2015). Development of a Self-Determination Measure for College Students: Validity Evidence for the Basic Needs Satisfaction at College Scale. *Measurement and Evaluation in Counseling and Development*, 48(4), 266-284. DOI: 10.1177/0748175615578737

Appendix 3. Frequency of requests addressed to experts

No.	Tasks *(in the general order of needs arising during the test development process)*	Type of expert input*	Total number of requests by test developers (see section 4)
1	defining/delimiting of the respective content area/domain	S	**2**
2	identifying domain specificity of items/tasks/problems	S	**10**
3	assessing the accuracy and appropriateness of translation/transfer of test items from one language/cultural context to another	S	**8**
4	judging/ranking items/tasks/problems along domain-/work-related criteria:		
	a) relevance (e.g., as facet of a complex competence) ……..	S	**20**
	b) importance/significance (e.g., for a profession)	S	**9**
	c) representativeness/typicality (e.g., a field of activity) ……...	S	**8**
	d) dangerousness (risk of harm or endangering others or damaging materials in completing a test item) ……...	S	–
	e) occurrence on the job (yes/no) ……	T_o	**2**
	f) frequency of occurrences on the job ……...	T_u	–
	g) content dimensions of the respective domain ……	T_o	**2**
5	judging/ranking/assigning items/tasks/problems according to ...…		
	a) test taker related criteria		
	aa) difficulty/complexity (for a certain group of test takers)	S	**4**
	ab) gender-specificity/-neutrality ...……...……..	T_u	–
	ac) cultural fairness ……...……..	S	**1**
	ad) aspiration level (high, medium, low; by scale) ……...……..	S	**2**
	b) psychological categories (e.g., classifying items to stages of cognitive, affective, psycho-motoric taxonomies)	T_o	–
6	evaluating the representativeness of a number of items as a sample drawn from the universe of a particular domain	T_u	**5**
7	ensuring the quality of wording/phrasing of items in terms of …		
	a) clarity (e.g., of phrasing) ……..…..	S	**15**
	b) understandability (e.g., foreign/abstract words) …….....	S	**6**
	c) grammar ……..…	T_o	**4**
	d) unambiguousness (uniqueness) …..……	S	**4**
	e) consistency (e.g., in use of terms) ……..….	T_o	**2**
	in the case of MC format:		
	f) correctness of attractors ……..…	T_o	**1**
	g) falsity of distractors ……	T_o	**1**
	h) plausibility of distractors ……	S	–
8	categorising items according to their …		
	a) feasibility (e.g., time constraints) ……...	S	**1**
	b) suitability/appropriateness (e.g., for a dynamic version) ...….	S	**4**
	c) curricular sequence in learning/educational process ……...	T_o	–
	d) logical sequence (e.g., solution independent of other items) ..	T_o	–
	e) psych. sequence (e.g., from easy/simple to difficult/complex)	S	–
	f) curricular significance (high, medium, low; by scale) ……	S	**2**
	g) opportunities to have been learned ……..	T_u	–

Note. * Cf. Table 1: T_o: truth-apt; T_u: truth-apt, but unknown; S: subjective