Assessing the Validity of Experts' Value Judgment over Research Instruments

ID No. 948

(PP 324 - 343)

https://doi.org/10.21271/zjhs.27.5.21

Rozhgar Jalal Khidhir
College of Basic Education,
Department of English, Salahaddin Univeristy-Erbil
rozhgar.khidhir@su.edu.krd

Tahsin Hussein Rassul
College of Basic Education,
Department of English, Salahaddin Univeristy-Erbil
tahsinhussein82@gmail.com

Received: 12/02/2023
Accepted: 27/03/2023
Published: 15/10/2023

Abstract

This study is primarily designed to assess the proportion of the usefulness, soundness and appropriateness of the experts' overall judgments over some developed instruments, such as interviews and questionnaires by researchers in the applied linguistics research area. The study has used a mixed method tool for data collection, such as a questionnaire and content analysis.

This study tries to answer the following questions: (i) In the researchers' viewpoint, how satisfactory is the experts' validation of the developed instruments?, (ii) to what degree do the developed instruments of researchers align with the academic and educational standards?, (iii) how do the experts' value judgments of the instruments align with the academic and educational standards?, (iv) what are the main challenges facing the researchers pre, during and post-validation process of the instruments?

The study has concluded the following points. The researcher participants, based on their perceptions, have provided vague impressions with regard to the experts' value judgment of the instrument validations, i.e., they remained uncertain of their efforts during instrument validation. The results of the content analysis have confirmed that the experts' value judgments hardly ever aligned with the academic and educational standards of validation. By contrast, it has been disclosed that the researchers have partially followed the academic and methodological standards for designing and developing any research instruments. Additionally, finding experts specialized in the psychometric domain has been considered an underlying obstacle during the validation of the instruments.

Keywords: Validity, Assessment, Judgment, Research Instrument.

1. Introduction

Research instruments are part and parcel of any study for data collection. So, studies need to have tools for gathering information about a phenomenon, concept, characteristic, etc. Although such instruments are of paramount importance of any research, there are often cases where poor data are collected owing to incomplete and inaccurate statements or questions, wording problems, and poor development process. These problems are serious and can be evaded or alleviated (Gillham, 2008).

Validity can ensure the quality of research instruments and mitigate the aforementioned critical issues relevant to 'newly constructed instruments'[1]. In validating any instrument, two problems commonly arise in KRI: First, researchers usually send their designed tools with insufficient information for the jury members to provide proper feedback. Second, jury members often provide various or inconsistent forms of feedback on checking validity of instruments. This paper seeks to highlight the missing information on researchers' designed

[1] Though there are situations where a new research instrument is required, using an existing instrument which has been used successful in previous research saves time and effort to design and validate. The current results can more easily be correlated with previous results (Price, et al., 2017).

instruments as well as to develop a standard guideline or template for validity reviewers to provide consistent and comprehensive feedback.

2. Review of Literature

The concept of validity, meaning a measuring tool is valid if it measures what it claims to measure, was first stated by Kelly (1972, P. 14 cited in (McLeod, 2013). To clarify the emphasis of the definition, Lakshmi & Mohideen (2013) indicate that the focus is not primarily on scores or items, but rather on inferences made from the instrument under investigation. For a research instrument to be considered valid, its inferences or interpretations ought to be "appropriate, meaningful, and useful" (Lakshmi & Mohideen, 2013, p. 2755).

Validity is considered one of the widely used quality criteria that any measuring instrument should meet so as to be employed by researchers in their studies (Fernandez-Gomez, et al., 2020). "A social science instrument measures latent variables that are not directly observed, although inferred from observable behaviour" (Bollen 2002, cited in (Elangovan & Sundaravel, 2021). Thus, social science instruments need to be verified whether they actually measure what they are intended to measure. Validity includes various types, namely *Face Validity*, *Construct Validity*, *Criterion-Related Validity*, and *Content Validity.*

Face Validity

It is "the extent to which a measurement method appears on its face to measure the construct of interest" (Price, et al., 2017, p. 70). Although regarded as the easiest and fastest type of validation, face validity is not sufficient because it is subjective (Elangovan & Sundaravel, 2021). Furthermore, Price, et al. (2017, p. 70) state that face validity "is a very weak kind of evidence" of validation because of two reasons: First, it is dependent on individuals' intuitions about human behaviour, which are often incorrect. Second, many psychometric instruments "work quite well despite lacking face validity". Additionally, McLeod (2013) believes that the term 'face validity'[2] should be avoided when the judgment is done by experts as 'content validity' in which the latter seems more appropriate.

Construct Validity

Nikolopoulou (2022) defines construct validity as ensuring how well a research instrument measures what it is supposed to measure. A construct is a concept or trait (such as, depression or job satisfaction) which is normally unobserved, but measured by observing other indicators that are connected to itTo establish construct validity for a measuring tool, it is required to ensure that a tool designed to measure a particular construct correlates with other tools assessing the same construct (i.e., convergent validity); and that two tools that should not be highly related to each other are actually unrelated (i.e., divergent or discriminant validity) (Bolarinwa, 2015; Nikolopoulou, 2022).

Criterion-Related Validity

This type of validity is adopted when there is an interest in correlating test results with another criterion of interest (Phelan & Wren, 2005). According to McLeod (2013) and Bolarinwa (2015), criterion-based validity pertains to how well the results of a measuring tool correlate with the currently existing criterion (i.e., concurrent validity) or to a future criterion (i.e., predictive validity.

Content Validity

Content validity is defined as "the degree to which elements of an assessment instrument are relevant to, and representative of, the targeted construct for a particular assessment purpose"

[2] Participants who are supposed to later respond to a research instrument are very likely to be considered suitable people to judge its face validity (McLeod, 2013).

(Haynes, et al., 1995, p. 238). In behavioural sciences, research and practice relevant to content validity is crucial for confirming sound measurement of research tools (Sireci, 1998).ddleton (2019), the content of a research tool (i.e., a test, questionnaire, and so forth) must include only and all relevant aspects of the concept it claims to measure. Elangovan & Sundaravel (2021, p. 3) confirm that content validity is about "domain relevance and representativeness of the test instrument". Additionally, Messick (1995) describes content validity in terms of the relevance of the content, representativeness and technical quality.

Expert Judgment

To validate the content of any instrument, the employed technique is fundamentally by consulting experts which is through a procedure called 'expert judgment'. Content validation via expert judgment is defined as "an informed opinion from individuals with a track record in the field who are regarded by others as qualified experts and who can provide information, evidence, judgements, and assessments" (Escobar-Pérez & Cuervo-Martínez, 2008 cited in (Fernandez-Gomez, et al., 2020, p. 2).

According to Lakshmi & Mohideen (2013, pp. 2755-6), experts suggest that four steps should be utilized to effectively evaluate content validity, namely *identifying and outlining the construct of interest*, *gathering information from domain experts*, *developing consistent matching methodology*, and *analysing results from the matching task*.

In addition, Hufford (2021) states that content validity can be achieved via sending measuring instruments to experts in the field where the relevant domain, items' relevance to the trait of the content, and the intended population need to be explained.

In their cooperative work, American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) established 'standards for educational and psychological testing' where it can be summarised as content validity needs to vividly include: *how the scores are aimed to be interpreted and consequently used; what method of sample selection is used for the population(s) indicating their socio-demographic and developmental characteristics; what construct(s) are under investigation in the instrument; describing the procedures for selecting judgments or ratings where the qualifications and experience of judges, any provided training and instructions, their agreement procedures should be presented; the allotted time for answering the tool; what prior preparation or motivation is needed for participants; the mode of instrument administration (e.g., online or face to face, invigilated or uninvigilated); the appropriateness of the instrument content;* and *accessibility of its content to all sample members* (AERA, et al., 2014).

In another study, Martinez believes that maximising the content validity of any instrument is via extensively reviewing literature and consulting with experts (2017, cited in (Elangovan & Sundaravel, 2021). According to Elangovan & Sundaravel (2021), those who are regarded as 'experts' include domain or subject matter experts, people with the expertise in designing tools, those taking decisions based on the instrument scores, and data analysts.

To validate a research instrument, experts are supposed to provide inputs on: how the definition of construct is related to the main domain in theory; the representativeness and significance of each item to the construct; accuracy of each item in measuring the concept; inclusion or deletion of elements; logical sequence of the items; scoring models; checking for bias; common errors such as double-barrelled, confusing, and leading questions; and a translated version of the instrument, if it is existing (Elangovan & Sundaravel, 2021, pp. 5-6).

According to Hinkin (1995), as part of developing instruments, experts should be provided with sufficient information with regard to the theoretical literature used for designing new instruments as well as the manner of item development.

Elangovan & Sundaravel (2021) state that despite of being provided with a qualitative account, reviewers or experts are often asked to rate instruments via using a rating form (i.e., quantitative assessment).

Besides, Bolarinwa (2015) describes how to achieve a content-valid instrument via focusing on several points, namely: *selecting experts familiar with the construct of interest or experts on the research subject*; *asking for reviewing the instrument items/questions for readability, clarity, and comprehensiveness; indicating the mechanisms of reaching an agreement rate in experts' item/ question ratings via using quantitative assessment called item-rating content validity indices (I-CVI) or scale-level rating (S-CVI).* A group of authors propose using the Content Validity Index (CVI) for quantitatively evaluating research instruments with a Likert scale (having options such as *Essential; Not Essential;* and *Modify)* where the accepted range of judges' agreement is at least 0.80 (Souza et al., 2017 cited in (Elangovan & Sundaravel, 2021).

To validate research instruments, both researchers and experts need to have mutual understanding of the required information for validation. The authors believe that the required information includes covering letter, introduction to research, construct-wise item validation, validation of demography items, and inferring the feedback. In other words, instrument developers should properly convey their requirements to the experts, and experts should know what aspect and howo validate any instrument (Elangovan & Sundaravel, 2021).

3. Methodology

This study is categorized as an evaluative endeavor that seeks for assessing expert reviewers' value judgments over research instruments through a mixed- method approach, i.e., quantitative and qualitative, data collection. The tools of the study consist of a structured closed and open-ended questionnaire and the content analysis for assessing the validation of the experts' value judgments. Beside, experts' value judgments are assessed twice; first, owing to the responses of the researchers and on the basis of several validity facets including finding experts, their approval, effort exertions, subjective judgment, and diligent revision process; second, due to the content analysis of the experts' revision records.

The subjects of the study fall into two folded parts, firstly, researches holding MA or PhD degrees and secondly, records of experts' reviews upon validating the research instruments.

The validity and reliability of both instruments have been achieved owing to a pilot study. Content and) face validity (2 were used to find out the validity of the instrument. As for the latter, internal consistency reliability using Cronback's Alpha was applied and the result was 0.885.

Different statistical tests, including Frequency and Descriptive Tests using SPSS program, have been applied to analyze the data.

4. DATA COLLECTION

Data Analysis of the Questionnaire

A. Part One: Demographic Information

The overall results from the descriptive statistics test –frequency for the participants' qualifications, shown in table 1 below, (**75%)** of the participants - 30 respondents, who responded to the questionnaire have MA degree, whereas only (**25%**) of them -10 respondents, are PhD holders.

Table 1. Researchers' Qualifications

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	MA	30	75.0	75.0	75.0
	PhD	10	25.0	25.0	100.0
	Total	40	100.0	100.0	

Likewise, the descriptive statistics test –frequency in table 2 below shows different outputs in terms of respondents' academic ranks as 45% assistant lectures in the highest position, and conversely, only 2.5% of the respondents as professors.

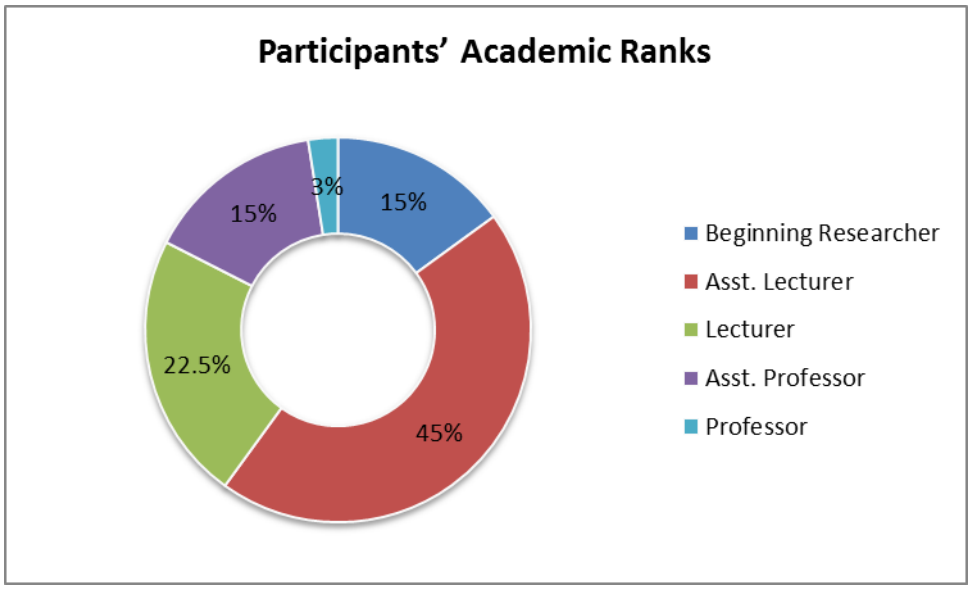
Table 2. Participants' Academic Ranks

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Beginning Researcher	6	15.0	15.0	15.0
	Asst. Lecturer	18	45.0	45.0	60.0
	Lecturer	9	22.5	22.5	82.5
	Asst. Professor	6	15.0	15.0	97.5
	Professor	1	2.5	2.5	100.0
	Total	40	100.0	100.0	

(2) **Expert information for the validity of the questionnaire**

No.	Name	Rank	University	Specialty
1	Dr. Ayad H. Mahmood	Professor	Diyala	Applied Linguistics
2	Dr. Fatima Rasheed	Professor	Salahaddin	Applied Linguistics
3	Dr. Himdad A. Muhammad	Professor	Salahaddin	Linguistics
4	Dr. Hussein A. Ahmed	Professor	Nawroz	Applied Linguistics
5	Dr. Dlakshan Y. Othman	Assistant Professor	Salahaddin	Applied Linguistics
6	Dr. Qismat M. Hussein	Assistant Professor	Salahaddin	Applied Linguistics

Chart 1. Participants' Academic Ranks



The descriptive statistics test revealed the mean score of (**6.27**) as the average number of studies done by the respondents. Meanwhile, the highest and the lowest number of the studies are (**60**) and (**0**) respectively. Table 3 below shows the detailed information about the average number of studies done by the respondents.

Table 3. The Mean Score of the Number of Researches

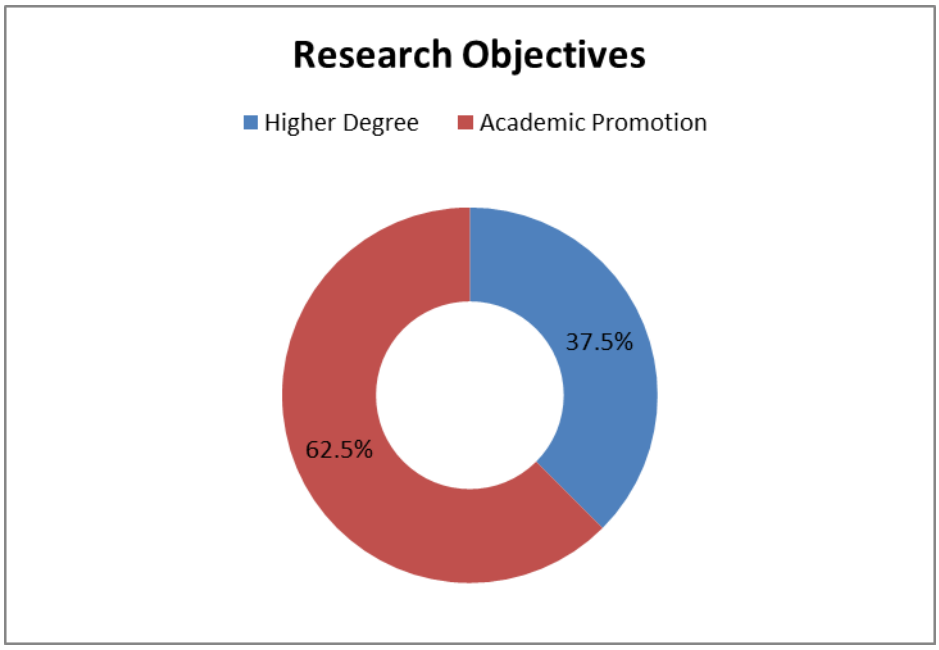
		NumberofResearch
N	Valid	40
	Missing	0
Mean		6.2750
Median		3.0000
Std. Deviation		9.94855
Minimum		.00
Maximum		60.00

Table 4 and chart 2 display the percentage value of the research objectives conducted by the respondents as **(37.5%)** for **higher degree** and (62.5%) for **academic promotion**.

Table 4. Research Objectives

		Frequency	Percent
Valid	Higher Degree	15	37.5
	Academic Promotion	25	62.5
	Total	40	100.0

Chart 2. Research Objectives



b. Part Two: Validation Evaluation

The descriptive statistics test for the Second Part of the questionnaire, as shown in table 6 below, reveals the following statistical outputs for each item separately: mean, score interpretation and standard deviation.

On the basis of the corresponding value scale range set to analyze and interpret the mean scores of each item on the questionnaire, the following outputs have been calculated. Only this set of items (**4, 5, 6, 7, 15, 16, 21 and 24)** has been introduced as (**S**) "**satisfied**" by the respondents due to the obtained mean score values starting from **3.50 to 4.20,** respectively. In the meantime, the other 20 items have been introduced as (**N**) "**neither satisfied nor dissatisfied**" **N**. Nonetheless, item (**6**) has received the highest mean score value (**3.92**) and (SD: **0.85**), by contrast, item (**9**) has received the lowest mean score value (**2.75**) (SD: 0.83). It is worth noting that the corresponding value scale ranges of mean score interpretations are as follows: VD=1.00-1.80, D= 1.90-2.60, N= 2.70-3.40. S= 3.50-4.20, VS= 4.30-5.00.

Table 5. Descriptive Statistics- Questionnaire, Validation Evaluation

	N	Mean	Score Interpretation	Std. Deviation
1	40	**3.3750**	**N**	.86787
2	40	**3.1000**	**N**	1.05733
3	40	**3.2500**	**N**	1.05612
4	40	**3.6000**	**S**	.95542
5	40	**3.5250**	**S**	.96044
6	40	**3.9250**	**S**	.85896
7	40	**3.6250**	**S**	.86787
8	40	**3.2500**	**N**	.80861
9	40	**2.7500**	**N**	.83972
10	40	**3.2750**	**N**	1.10911
11	40	**3.4000**	**N**	.95542
12	40	**2.9250**	**N**	1.02250
13	40	**3.1500**	**N**	.92126
14	40	**2.9000**	**N**	.95542
15	40	**3.5000**	**S**	.84732
16	40	**3.5500**	**S**	.84580
17	40	**3.1500**	**N**	.92126
18	40	**3.3500**	**N**	.92126
19	40	**3.4750**	**N**	.87669
20	40	**3.4750**	**N**	.81610
21	40	**3.5250**	**S**	.67889
22	40	**3.3250**	**N**	.88831
23	40	**3.3250**	**N**	.82858
24	40	**3.6750**	**S**	.82858
25	40	**3.3750**	**N**	.97895
26	40	**3.4250**	**N**	.81296
27	40	**3.4750**	**N**	.93336
28	40	**3.3500**	**N**	.83359

c. Part Three: Instrument Evaluation

Descriptive statistics test for the **third part** of the questionnaire revealed the following mean score outputs. As shown in table 7 below, **12 out of 23** items on the questionnaire have been ranked between a range scale starting from **1.80** to **2.49** which has been introduced as "**partially**", whereas the rest of other responses were revealed as "**entirely**" on the basis of their mean score interpretations. This result indicates the sole average of the respondents' responses to this part. Nevertheless, item (**2.3**) has received the highest mean score value (**2.82**) and (**SD**: **0.38**), by contrast, item (**2.20)** has received the lowest mean score value (**1.82**) and (**SD: 0. .81**). It is worth noting that the corresponding value scale ranges of mean score interpretations are as follows: not at all =1.00 -1.79, partially= 1.80-2.49, entirely= 2.50-3.00.

Table 6. Descriptive Statistics- Questionnaire/ Instrument Evaluation

	N	Mean	Score Interpretation	Std. Deviation
1	40	**2.4000**	**Partially**	.67178
2	40	**2.5500**	**Entirely**	.55238
3	40	**2.8250**	**Entirely**	.38481
4	40	**2.6000**	**Entirely**	.54538
5	40	**2.8000**	**Entirely**	.51640
6	40	**2.6250**	**Entirely**	.58562
7	40	**2.4250**	**Partially**	.71208
8	40	**2.0500**	**Partially**	.74936
9	40	**2.2500**	**Partially**	.70711
10	40	**2.6500**	**Entirely**	.57957
11	40	**2.7500**	**Entirely**	.49355
12	40	**2.6000**	**Entirely**	.59052
13	40	**2.5250**	**Entirely**	.64001
14	40	**2.4500**	**Partially**	.59700
15	40	**2.5250**	**Entirely**	.59861
16	40	**2.6500**	**Entirely**	.48305
17	40	**2.2250**	**Partially**	.73336
18	40	**2.4250**	**Partially**	.63599
19	40	**1.9000**	**Partially**	.77790
20	40	**1.8250**	**Partially**	.81296
21	40	**2.4000**	**Partially**	.63246
22	40	**2.1000**	**Partially**	.63246
23	40	**2.3000**	**Partially**	.64847

d. Part Four: Validation Challenges

Descriptive statistics test for **the fourth part** of the questionnaire revealed the following mean score outputs: **11** out of **20** items on the questionnaire have positioned between a range scale starting from **1.80** to **2.49** which has been introduced as "**not sure**", whereas; the other responses were revealed as "**disagree (8 items)**, except for item **2** which has been estimated as **agree** based on their mean score interpretations. This result indicates the sole average of the respondents' responses to this part. Nonetheless, item (**2**) has received the highest mean score value (**2.50**) and (SD: **0.59**) by contrast, item (**9)** has received the lowest mean score value (**1.32**) and (**SD: 0.57).** It is worth noting that the corresponding value scale ranges of mean score interpretations are as follows: disagree=1.00-1.79, unsure=1.80-2.49, agree= 2.50-3.00.

Table 7. Descriptive Statistics- Questionnaire/ Validation Challenges

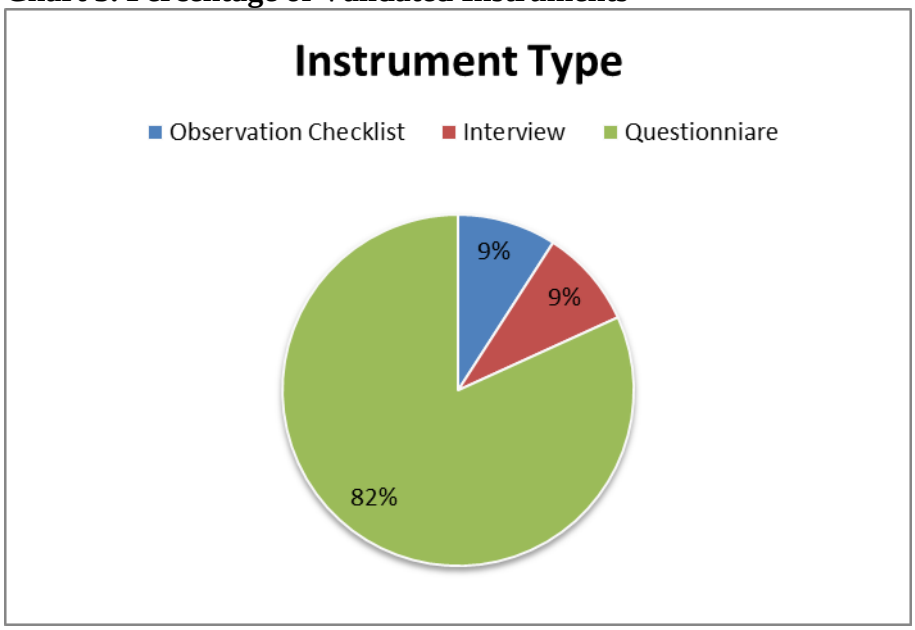
	N	Mean	Score Interpretation	Std. Deviation
1	40	**2.1500**	**Not Sure**	.80224
2	40	**2.5000**	**Agree**	.59914
3	40	**1.5500**	**Disagree**	.78283
4	40	**1.7250**	**Disagree**	.81610
5	40	**1.4250**	**Disagree**	.74722
6	40	**1.5250**	**Disagree**	.78406
7	40	**1.8250**	**Not Sure**	.74722
8	40	**1.4250**	**Disagree**	.67511
9	40	**1.3250**	**Disagree**	.57233
10	40	**1.9000**	**Not Sure**	.77790
11	40	**1.7750**	**Disagree**	.89120
12	40	**2.1750**	**Not Sure**	.71208
13	40	**1.4000**	**Disagree**	.77790
14	40	**1.9750**	**Not Sure**	.76753
15	40	**2.0250**	**Not Sure**	.83166
16	40	**2.1750**	**Not Sure**	.78078
17	40	**2.4500**	**Not Sure**	.74936
18	40	**2.0750**	**Not Sure**	.69384
19	40	**2.3250**	**Not Sure**	.69384
20	40	**2.1000**	**Not Sure**	.74421

Data Analysis: The Document Content Analysis

On the basis of the history of recorded responses (notes and comments) of the experts during the validation of the instruments, the following results have been recorded.

The overall analyzed records of the experts are (44) documents categorized as questionnaires, interviews and checklists. The percentage of each validated instrument type by the experts below is depicted in Chart 3.

Chart 3. Percentage of Validated Instruments



Furthermore, the records have been analyzed and assessed twice by different raters- assessors, first by the researchers and then by another invited co-rater [(1)]. The below table shows the statistical test output for inter-rater reliability of different scoring values. In fact, all the three raters are specialized in the content area and the psychometric domain, respectively. Besides, it is worth stating that the whole records of the experts have been analyzed and assessed by the raters at the physical aspects' par, that is, in terms of evident recorded feedback such as brief remarks, comments, short messages, digits, ticks and any other indications.

Table 8. Inter-Rater Reliability

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of Agreement	Kappa	.540	.082	10.576	.000
N of Valid Cases		44			

a. Not assuming the null hypothesis.

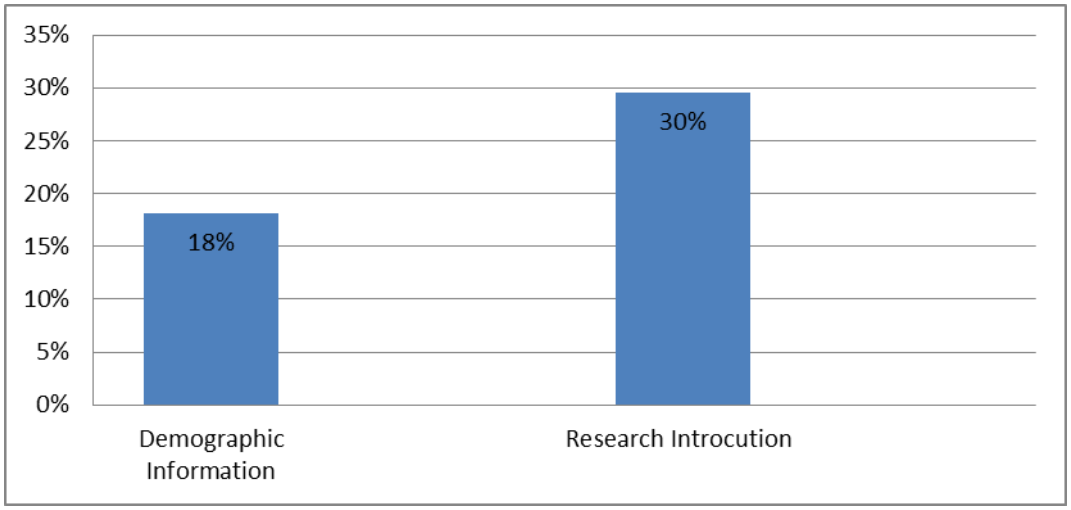
b. Using the asymptotic standard error assuming the null hypothesis.

Notably and on the other hand, the content analysis of the history records of the documents, i.e., instruments, has gone through twofold accurate analysis: First and foremost, overall items of the content – each based on certain criteria (15 scales), have been analyzed individually, then, for the ease of score interpretation, several specific themes were formed to which each individual item was annexed following the process of theme formation. The themes have been mentioned in the titles of the following charts alternately.

As shown in Chart 4, the percentage value reveals that only 18% of the experts reviewed and commented on demographic information on the instrument. Likewise, 30% of them reviewed and commented on the research design and description.

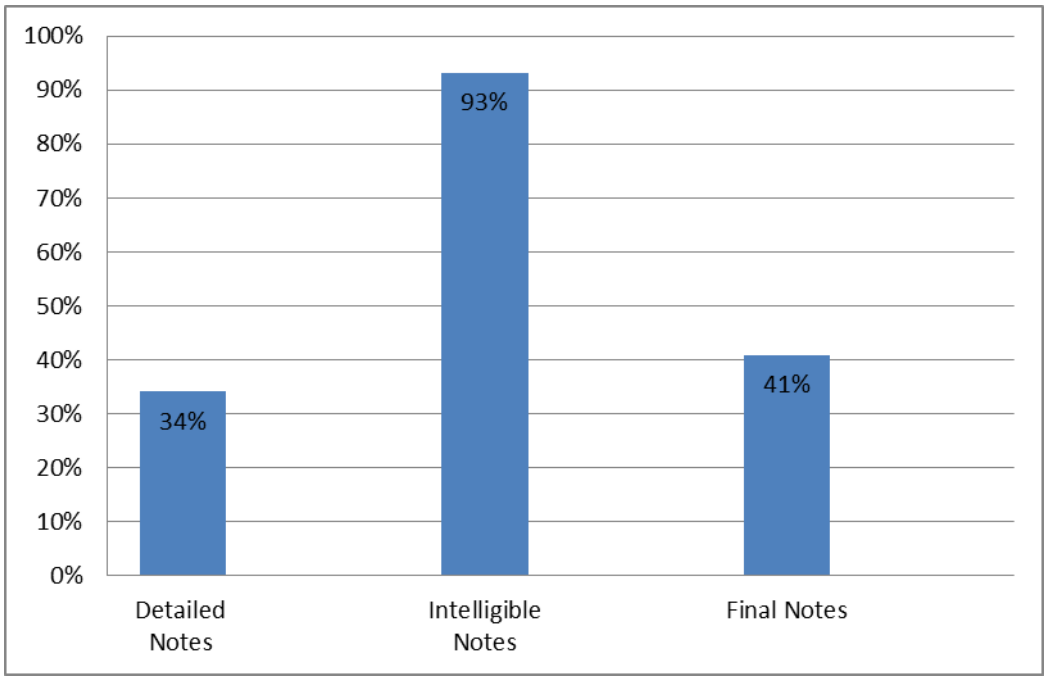
(1) Dilakhshan Y. Othman, PhD, Assistant Professor, University of Salahaddin.

Chart 4. Review of Instrument Introduction



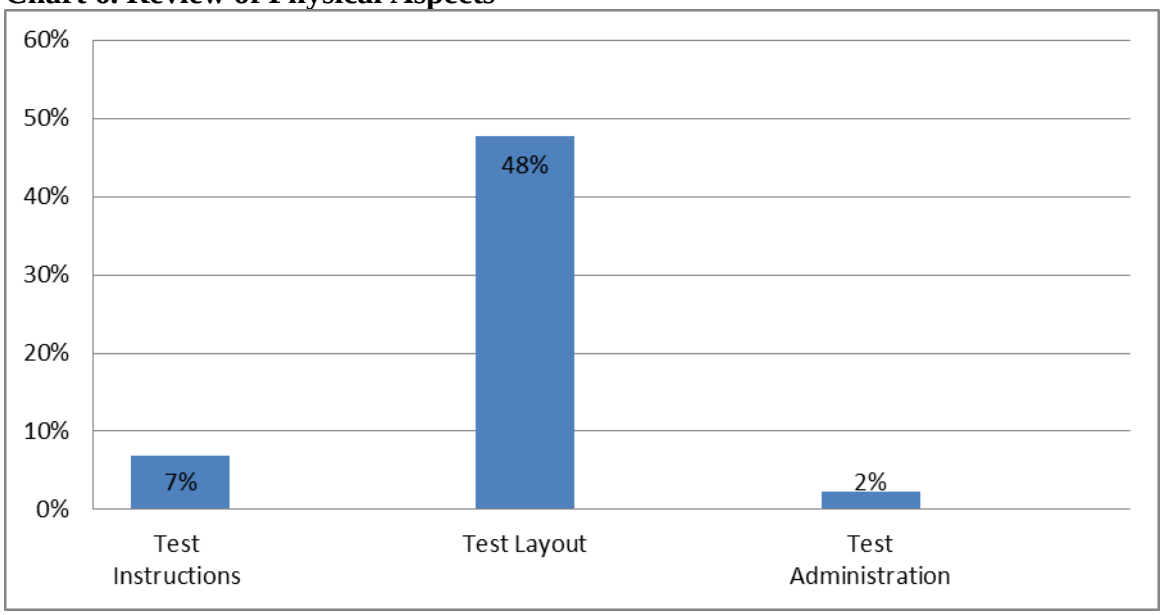
The percentage rates in Chart 6 for commenting and giving notes, reveal the following outputs: (34%) of the experts provided detailed notes on the instruments they reviewed. Besides, (93%) of the notes the experts provided were intelligible. However, only (41%) of the notes has been categorized as final notes.

Chart 5. Commenting and Giving Notes



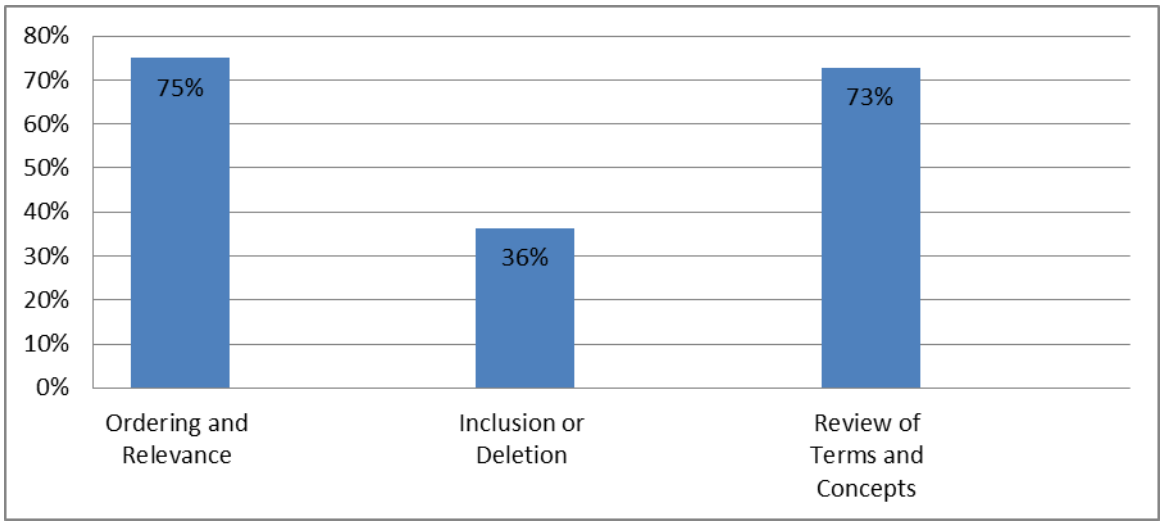
The percentage rates as shown in Chart 7, for the review of physical aspects, reveals the following outputs. Only 7% of the experts reviewed and commented on the instrument instructions. Meanwhile, **48%** of them reviewed and commented on the layout and design of the instruments. Last but not least, 2% of them reviewed and commented on test administration.

Chart 6. Review of Physical Aspects



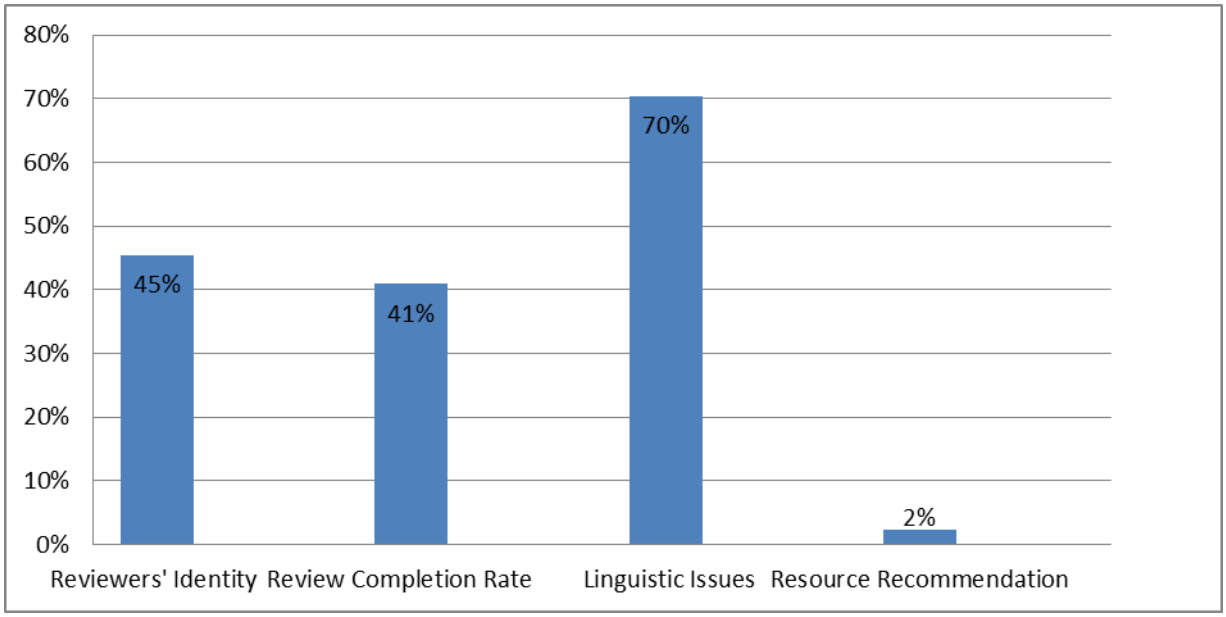
The frequency and percentage test, chart 8 for review of items, reveals the following outputs. Significantly, **75%** of the experts reviewed and commented on the ordering and relevance of every single item on the test. Moreover, only **36%** of them reviewed and came up with suggestions for adding up and/or deleting specific items. Lastly, **73%** of them reviewed and commented on different terms and concepts used throughout the test.

Chart 7. Review of Items



The percentage rate, shown in Chart 9, for the review of other related points, reveals the following outputs. In fact, **45%** of the experts provided various information associated with their identity and academic profile. Additionally, **41%** of them reviewed the test thoroughly. Notably, **70%** of the experts reviewed and commented on the linguistic aspects of the instruments. Finally, only **2%** of them recommended extra resources.

Chart 8. Review of Other Related Points



5. Interpretation of Results

a. Questionnaire

Multiple interpretations are inferred on the basis of the obtained results from the questionnaire. Hence, they are posed in accordance with the order of the parts and items on the questionnaire.

With regard to demographic information, the number of the participants with MA degrees is three times bigger than those with PhD degrees (as shown in Table 1). This alludes that researchers with MA degree are yet want to make progress and more keep vibrant in doing researches than PhD ones. This is most likely due to the intention to either pursuing higher qualifications or academic ranks.

Additionally participants with assistant lecturers and lecturers as academic titles are amongst the highest numbers of the participants responded to the questionnaire. This is possibly owing to the intention of making further progress in their academic profile and career. Yet, as they are at the very outset of research doing phase, they are most likely to face difficulties in coping with research methodologies and their principles, especially in designing and developing research instruments.

As for the second part of the questionnaire, the average mean scores between **2.70 to 3.40** indicate that the respondents seemed skeptical, i.e., unsure, about the overall quality of the validation of the experts. Most importantly, the deduced findings reflect the respondents' views concerning this interpretation. Thus, the first research question as, ***"In the researchers' viewpoints, how satisfactory is the experts' validation of the developed instruments?"*** is answered.

With regard to the third part of the questionnaire, the obtained responses typically show that the developed tools of the participants **partially** align with the academic and educational standards. That is, this inferred finding is very likely to bear different interpretations including: Firstly, it seems that the researchers are not fully aware of the academic and methodological standards of instrument design and development, secondly; they may not have taken adequate courses concerning research methodology and its guidelines.

Besides, the deduced findings stemmed from the respondents' perceptions that they have been quite fair with their judgments over their own performance in test development process. Thus, the second research question as ***"To what degree do the developed instruments of researchers align with the academic and educational standards?"*** is responded.

On the basis of the obtained results from the fourth part of the questionnaire, it can be inferred that the researchers on average are uncertain concerning the challenges they encounter pre, during and post the validation of the instruments (See Table 8 for more details). Meanwhile, the respondents did not add any other barriers to the rest in the last open-ended item on the questionnaire. However, most researchers sided with item 2: "I cannot find a sufficient number of experts specialized in the psychometric domain." In other words, finding sufficient number of experts in psychometric domain is said to be the most common issue amongst researchers. Thus, the third research question, "***What are the main challenges facing the researchers in pre, during and post-validation processes of the instruments?"*** is answered.

b. Documents

The interpretations of the data of the content analysis of the experts' records give rise to several observed insights which can be briefly reported as follows.

Review of Instrument Introduction

A small number of experts have reviewed and commented on the demographic information questions on the developed instruments (See Chart 5 for extra information). The perceived finding causes some disappointment concerning validation of various instruments since the experts seemed to have paid no or very little attention to them. However, one more interpretation might be posed that there tends to be a lack of awareness by the experts to review the demographic information as well, even though they are not invited for. In the meantime, much less than half of the experts representing 30% similarly reviewed and commented on the research introduction, such as the title of the study, research questions, aims, hypotheses, and etc. Again, this result infers that many of the experts are likely to have been either careless about this point or not specialized in the content area of the subject and the construct of the reviewed instruments.

Review via Commenting and Giving notes

In fact, much less than half of the reviewers representing 34%, failed to provide detailed notes on the reviewed tools. Meanwhile, almost half of them, 41%, left final notes at the end of the tools for the researchers to take into account. In contrast, almost all the proportional notes, 93%, provided by the experts were clear and understandable. Thus, one can comprehend that the experts have not been serious or careful about giving detailed notes and final remarks as feedback to the researchers to improve their developed tools. Conversely, the experts did very well in providing intelligible notes as a whole.

Review of Physical Aspects

This aspect of validation has received no or the least attention from the experts since the obtained result clearly shows. Only 7% of the experts reviewed and commented on the test instructions of the instruments. Likewise, only 2% of them reviewed and commented on the test administration including the policy, time, place and the characteristics of the participant. These findings allow a dual interpretation: First, it is most likely that the experts were unaware of taking these points into account during validation, secondly, the test developers may not have clearly exhibited these points on the test. Even if the latter claim appears to be true; yet, the experts should have warned the test developers about it, at least in a comment.

On the other hand, almost half of them, 48%, reviewed and commented on the test layout such as clarity, simplicity, practicality and so on. The concluded percentage is not likely to be satisfying or encouraging by any research instrument developers.

Review of Items

This part of the questionnaire has gained most of the attention by the experts acting upon the obtained results. Approximately, three quarters of them reviewed and commented on the ordering scheme of the items, their relevance to the construct and the special terms and concepts utilized in the instrument. Despite this, their reviews and comments sporadically varied in terms of continuous occurrence and length. Besides, efforts by the experts for the inclusion and deletion of the items were sometimes noted.

Review of Other Related Points

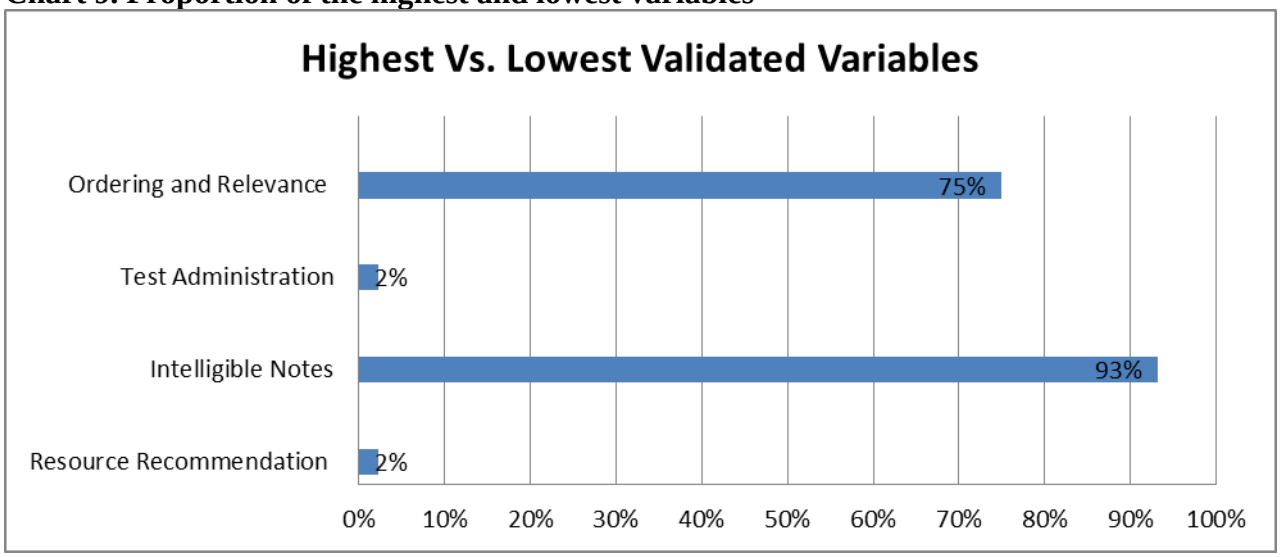
Under review of other related points, experts' (reviewers) identities, review completion rate, linguistic issues and resource recommendations are covered.

Significantly, nearly half of the reviewers provided information in association with their academic profile such as names, ranks, specialty and affiliation. Moreover, providing information about one's personal and academic identity is worth a lot in academic assessment and evaluation since it allows for authentic, objective and unbiased performance. Secondly, slightly less than half of the reviewers have reached the completion rate limit of validation of the tools they confirmed. Probably, this is partly due to the length and the inclusive content of the tools the developers provided and/or the reckless revision of the reviewers themselves. Thirdly, almost three quarters of them reviewed the linguistics features in terms of grammatical errors, misspellings, punctuation and capitalization mistakes, and misuses of vocabularies and sentence deep structure. Finally, extremely very few of the reviewers representing merely 2% recommended extra resources. Thus, it appears that the reviewers either did not consult any other resources for the validation or they were not specialized in the content area of the research instruments.

6. Summary of Findings

Generally, minor efforts have been exerted by experts in reviewing or providing adequate information about 11 aspects namely; demographic information, research introduction, academic identities, review completion rate, resource recommendation, providing full notes, final notes, test instruction, test layout, test administration, and adding or deleting items. By contrast, they performed well on reviewing and providing adequate information about 4 areas (i.e., clear feedback – notes, item relevance and order, terms and concepts, and linguistics issues). In the meantime, the lowest revision score has gone for two items as equal, including test administration and resource recommendation, likewise; the highest revision score has gone for providing intelligible notes. The chart below shows the summary of the items receiving the highest and lowest score, respectively.

Chart 9. Proportion of the highest and lowest variables



On the basis of both data analysis and interpretation of the results, it is imperative to claim that the experts, as already confirmed through their document history, typically failed to review the content of the research instruments of the researchers during validation. In other words, their value judgments poorly align with the academic and educational standards. Thus far, the fourth research question as **"*How do the experts' value judgments of the instruments align with the academic and educational standards?"*** is answered.

Significantly, the above-mentioned claim can be excused for the availability of several other potential justifications, including, first ; the tool developers, i.e., researchers, are most likely to be one of the major factors beyond such a failure of successful validation since they are not fully aware of the academic and methodological standards of instrument design and development, (See Table 7 for more descriptive details in this regard). Second, the reviewers seem to be either not specialized in the content areas of the developed tools or they might have validated the tools carelessly.

To sum up, overall indications of the inferred findings from this study tools suggest forming a standardized guideline for both researchers upon designing and developing tools- questionnaires, interviews, and tests; and expert reviewers upon tool validation.

7. Conclusion

Typically, the researcher participants, based on their perceptions, have provided vague impressions with regard to the experts' value judgment of the instrument validations, i.e., they remained uncertain of the experts' validation, as a result, the researchers' both performance and achievement of instrument validation will be in question.

Notably, the above point has been reiterated by the results of the content analysis as well. Equally, the content analysis results support such an aforementioned claim since their value judgments hardly align with the academic and educational standards.

On the other hand, the researchers, depending on their self-assessment, have partially followed the academic and methodological standards for designing and developing their research instruments. In addition, the researchers' failure to completely comply with the terms and conditions of instrument development can therefore bring about incomplete validation and thus be deemed as one of the underlying factors beyond the validation issues.

Additionally, the researcher participants have unanimously agreed upon considering finding experts specialized in a psychometric domain as an underlying obstacle during the validation of the instruments.

Thus, neither researcher participants nor expert reviewers participated in this study managed to follow a standard guideline to develop and validate instruments due to a couple of potential factors (inferred from the results): First of all, researcher participants are incompetently skilled or novice in research area since they failed to provide and convey the required information to the experts, , the unavailability of unified forms and guidelines concerning designing tools and their validation for research purposes.

8. Recommendations

Based on the conclusions, the study suggests forming a standardized guideline for both researchers upon designing and developing tools- questionnaires and interviews and expert reviewers upon tool validation as follows.

Recommended Guidelines for Instrument Developers and Validators

1. Instrument Developers

No.	Item	Description	Inclusion Option
1	**Formal Invitation**	Writing a formal letter requesting experts to review the instrument for validity	Mandatory
2	**Research Description**	Introducing the research briefly for which the instrument is specifically developed by including the background information, research questions, aims, significance, delimitations and limitations.	Highly Recommended
3	**Test (Instrument) Specification: Intended use of the instrument**	Setting forth the intended use/ uses of the instrument, e.g. for research purpose, by providing detailed information	Recommended
4	**Test (Instrument) specification: Intended**	Disclosing the identities of the users of the test such as researchers, examiners, syllabus	Recommended

	users of the test	designers, educational policy makers, etc.	
5	**Demographic Questions**	Asking the respondents to provide information in regard to their identities through closed or open-ended items. NB: The number and the type of the items vary according to the nature and the purpose of the study	Mandatory
6	**Population and Sample Size**	Disclosing the identities of the overall population of the study as well as the true number of the respondents representing and targeted to participate in the study	Mandatory
7	**Score Interpretations**	Interpreting the obtained results by comparing to predetermined scale by the researcher, i.e., norm-referenced or criterion-referenced- a validated scale resulted in previous researches.	Recommended
8	**Administration**	Including time and place, the date and the exact time of test taking should be mentioned as well as the venue of test administration such as full address, city, quarter, building, hall number etc.	Highly recommended
9	**Defining Variables**	The variables including dependent and independent ones should be clearly defined and referred to in the tool.	Recommended
10	**Layout**	The instrument layout in terms of the structure, content and look should be clearly sound and easy to follow.	Highly recommended
11	**Delivery Mode**	The means of delivering the test should be mentioned (e.g. internet- telephone- face to face)	Recommended
12	**Validity Type**	As there are various types of instrument validity, it is imperative that the instrument developers specify and highlight the type of validity they wish experts to validate the tool for them such as face, content, construct, etc.	Mandatory
13	**Instruction**	Instructions, such as directions and requirements of the tool, should be designed in such a way that they can be easily followed by both its participants and the experts for validation.	Highly recommended
14	**Experts' Academic Profile Information**	A space should be provided for the experts' academic profile upon validation completion for history records and validation creditability	Mandatory

2. Instrument Validators (Experts)

No.	Item	Description
1	Demographic questions	Their relevance to the research aims and questions
2	Research introduction	Research questions, aims and hypotheses in association with the construct
3	Language issues	double-barreled (overloaded items)
		double negative items
		confusing items
		leading questions (leading the respondents to answer)
4	Additional resources	Recommending further resources for instrument developers
5	Instrument instructions and directions	Short and clear instructions that are easily followed
6	Instrument layout	Clarity, simplicity and practicality of the developed instrument
7	Test administration policy	(time and place of test)
8	Items revision	accuracy of the use of terms and concepts

		item relevance and ordering
		inclusion or deletion of items
		sensitivity (e.g. personal life)
		bias

9. References

A. Literature Review

American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME), (2014). Washington, DC: American Educational Research Association.

Bolarinwa, O. A., (2015). Principles and methods of validity and reliability testing of questionnaires used in social and health science researches. *Nigerian Postgraduate Medical Journal*, 22(4), pp. 195-201.

Elangovan, N. & Sundaravel, E., (2021) Method of preparing a document for survey instrument validation by experts. *MethodsX*, 8, pp. 1-9.

Fernandez-Gomez, E. et al., (2020) Content validation through expert judgement of an instrument on the nutritional knowledge, beliefs, and habits of pregnant women. *Nutrients*, 12(4), pp. 1-13.

Gillham, B. (2008) *Developing a Questionnaire*. 2nd ed. London: Continuum International Publishing Group.

Haynes, S., Richard, D. & Kubany, E. (1995) Content validity in psychological assessment: A functional approach to concepts and methods. *Psychol Assess*, Volume 7, pp. 238-247.

Hinkin, T. R. (1995) A review of scale development practices in the study of organizations. *J. Manag.*, 21(5), pp. 967-988.

Hufford, B. (2021) *The 4 Types of Validity in Research Design* (+3 More to Consider). [Online] Available at: https://www.activecampaign.com/blog/validity-in-research-design [Accessed 3 December 2022].

Lakshmi, S. & Mohideen, M. A. (2013) Issues in reliability and validity of research. IJMRR, 3(4), pp. 2752-2758.

McLeod, S. A. (2013) What is Validity? *Simply Psychology*. [Online] Available at: www.simplypsychology.org/validity.html [Accessed 29 November 2022].

Messick, S. (1995) Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), pp. 741-749.

Middleton, F. (2019) *The 4 Types of Validity in Research*. [Online] Available at: https://www.scribbr.com/methodology/types-of-validity/ [Accessed 3 December 2022].

Nikolopoulou, K. (2022) *What is Content Validity*. [Online] Available at: https://www.scribbr.com/methodology/content-validity/ [Accessed 3 December 2022].

Phelan, C. & Wren, J. (2005) *Reliability and Validity*. [Online] Available at: https://chfasoa.uni.edu/creatinggoalsandoutcomes.htm [Accessed 3 December 2022].

Price, P. C. et al.(2017) *Research Methods in Psychology*. 3rd ed. the USA: s.n.

Sireci, S. G. (1998) The construct of content validity. *JSTOR*, 45(1), pp. 83-117.

B. Questionnaire and Guideline References

American Educational and Psychological Research Association (2014). Standards for Educational and Psychological Testing. New York: American Educational Research Association, the American Psychological Association,

Collingridge, D. (2021) *Validating a Questionnaire: Data Collection*. [Online] Available at: https://www.methodspace.com/blog/validating-a-questionnaire [Accessed 22 November 2022].

Ikart1, E.M. (2019) Survey Questionnaire Survey Pretesting Method: An Evaluation of Survey Questionnaire via Expert Reviews Technique. [PDF] *Journal of Social Science Studies* Vol. 4, No. 2. Available at: https://doi.org/10.20849/ajsss.v4i2.565 [Accessed 22 November 2022].

Kalkbrenner, Michael T. (2021) A Practical Guide to Instrument Development and Score Validation in the Social Sciences: The Measure Approach,, *Practical Assessment, Research, and Evaluation:* Vol. 26 , Available at: https://scholarworks.umass.edu/pare/vol26/iss1/1 [Accessed 22 November 2022].

Siegle, D. (2022) Instrument Validity. NEAG SCHOOL OF EDUCATION Educational Research Basics. [Online] Available at: https://researchbasics.education.uconn.edu/instrument_validity/#

Sundarave , E. & Elangovan , E. (2021) *Method of preparing a document for survey instrument validation by experts*. [Online]. Available at: http://creativecommons.org/licenses/by-nc-nd/4.0/ [Accessed 22 November 2022].

Tress, F. (2016) *Five Basic Principles for Writing Good Questionnaires*. [Online] Available at: https://norstatgroup.com/blog/five-basic-principles-for-writing-good-questionnaires [Accessed 22 November 2022].

هەڵسەنگاندنی دروستی بریاردانی پسپۆڕان لەسەر ئامرازە کانی توێژینەوە

رۆژگار جلال خضر
کۆلیژی پەروەردەی بنەڕەتی- بەشی زمانی ئینگلیزی- زانکۆی سەلاحەدین -هەولێر
Email: rozhgar.khidhir@su.edu.krd

تحسین حسین رسول
کۆلیژی پەروەردەی بنەڕەتی- بەشی زمانی ئینگلیزی- زانکۆی سەلاحەدین -هەولێر
Email: tahsinhussein82@gmail.com

پوختە

توێژینەوەیەکە بریتی یە لە هەوڵێك بۆ هەڵسەنگاندنی ڕێژەی دروستی و گرنگی و و گونجاوی بریاردانی گشتی شارەزایان لەمەر هەندێک لە ئامرازی ئامادەکراوی تویژەران وەک چاوپێکەوتن و ڕاپرسی لە بواری توێژینەوەی زمانەوانی کارەکی دا. توێژینەوەکە ئامرازێکی تێك هەلکێشی نێوان دوو جۆری وەك چەندێتی و چۆنیەتی بەکارهێناوە بۆ مەبەستی کۆکردنەوەی زانیارییەکان لەوانە ڕاپرسی و شیکاری ناوەڕۆک. لە دەرئەنجامەکانی توێژینەوەکە هاتووە : بەشداربووان، بە پشتبەستن بە تێڕوانینەکانیان، وەلامی نارووونیان پێشکەش کردووە سەبارەت پرۆسەو سەنگی بریاردانی شارەزایان لەبارەی دروستی و گونجاوی ئامرازەکانەوة ، واتە، توێژەران لە هەوڵەکانی شارەزایان لە کاتی پیداچونەوە بە ئامرازەکان دڵنیا نین، ئەنجامی شیکاری ناوەڕۆک پشتڕاستی کردووەتەوە کە بریاری پسپۆڕان لەمەر ئامرازەکانی توێژەران لەگەڵ ستانداردە ئەکادیمی و پەروەردەییەکانی دروستی و گونجاوی هاوتەریب نین، لەبەرامبەردا، ئەمە ڕوون بۆتەوە کە بەشێکی دیاریکراوی توێژەران ستانداردە ئەکادیمی و میتۆدۆلۆژییەکانیان بۆ دیزاینکردن و پەرەپێدانی ئامرازەکانی توێژینەوە پەیڕەو نەکردووە. سەرەڕای ئەمانەش، دۆزینەوەی شارەزایانی پسپۆڕ لە بواری سایکۆمیتریدا بە بەربەستێکی بنەڕەتی هەژمار کراوە.

وشەی سەرەکی: دروستی ، هەڵسەنگاندن، بڕیار، ئامرازی توێژینەوە

تقييم صحة أحكام الخبراء على أدوات البحث

رۆژگار جلال خضر
كلية التربية الاساسية - قسم اللغة الإنجليزية - جامعة صلاح الدين-اربيل
Email: rozhgar.khidhir@su.edu.krd

تحسین حسین رسول
كلية التربية الاساسية - قسم اللغة الإنجليزية - جامعة صلاح الدين-اربيل
Email: tahsinhussein82@gmail.com

الملخص

البحثُ عبارة عن محاولةٍ لتقييمِ نسبةِ صحةِ وأهميةِ ملاءمة القراراتِ العامةِ للخبراءِ حول بعضِ الأدواتِ المعدَّةِ من قبلِ الباحثين، مثلُ المقابلة والاستبيان في مجالِ البحوثِ اللغويةِ التطبيقيةِ. وقد استخدمَ البحثُ طريقةَ المُزاوجةِ بينَ نوعينِ من الأساليبِ هما الكيفيُّ والكميُّ، وذلك لجمعِ المعلوماتِ، من ذلك الاستبيانُ وتحليلُ المضمونِ. وقد جاءَ من نتائجِ الدراسةِ الآتي: المشاركون، وبالاعتمادِ على توجهاتهم، قدّموا أجوبةً غيرَ واضحةٍ حول العمليةِ وقيمةِ قرارِ الخُبراء تُجاهَ صحةِ وملاءمةِ الأدواتِ، وهذا يعني أنّ الباحثين ليسوا مطمئنين من مراجعةِ الخُبراءِ لتلكَ الأدواتِ؛ وقد أثبَتَتْ نتائجُ تحليلِ المضامينِ أنَّ قرارَ الخُبراءِ تُجاه أداوتِ الباحثين لا يتطابقُ مع المعاييرِ الأكاديميةِ والتربويةِ الصحيحةِ، في مقابلِ ذلك تبيَّنَ أنّ قسماً محدّداً من الباحثين لم يطبِّقوا المعاييرَ الأكاديمية والنظرية للتصميمِ وتطويرِ أدواتِ البحثِ، مع ذلك فإنَّ إيجادَ خُبراء في مجالِ الاختبارِ النفسيّ (السايكومتري) يُعَدُّ عائِقاً جذرياً.

الكلمة الرئيسية: صحة، تقييم، قرار، أداة بحث