LAPORAN PERTANGGUNGJAWABAN PELAKSANAAN KEGIATAN *VISITING PROFESSOR* DARI PROGRAM STUDI PENELITIAN DAN EVALUASI PENDIDIKAN PROGRAM PASCARJANA UNIVERSITAS NEGERI YOGYAKARTA (UNY) KE *FACULTY EDUCATIONAL STUDIES* UNIVERSITI PUTRA MALAYSIA TAHUN ANGGARAN 2019/2020

Unit Kerja : Program Pascasarjana (PPs) Universitas Negeri Yogyakarta (UNY)
Program : Pengembangan Kerjasama Program Studi Penelitian dan Evaluasi Pendidikan (PEP) PPs UNY dengan PT LN
Kegiatan : *Visiting Professor* untuk Prof. Dr. Badrun Kartowagiran, M.Pd. Program Studi Penelitian dan Evaluasi Pendidikan (PEP) Program Pascasarjana (PPs) Universitas Negeri Yogyakarta (UNY) ke Universiti Putra Malaysaia (UPM)

A. Analisis Pencapaian Kinerja Kegiatan	Kegiatan *Visiting Professor* yang dilaksanakan program studi PEP didasarkan pada Permenristekdikti No. 44 Tahun 2015 tentang Standar Nasional Pendidikan Tinggi dan Tujuan Renstra UNY Tahun 2015-2019 No 1, 2, dan 3. Selanjutnya, kegiatan ini dilatarbelakangi oleh rangkaian upaya UNY dalam mencapai *World Class University* (WCU), yaitu dengan melakukan kerjasama Perguruan Tinggi Luar Negeri (PT LN). Kegiatan *Visiting Professor* ini juga sesuai dengan visi dan misi UNY tahun 2045 untuk menjadi perguruan tinggi unggul di ASEAN. Kegiatan *Visiting Professor* ini dilakukan dengan mendelegasikan salah satu dosen program studi PEP untuk melaksanakan serangkaian aktivitas di universitas mitra. Dalam hal ini Prof. Dr. Badrun Kartowagiran merupakan pelaksna *Visiting Professor* yang dilaksanakan di Universiti Putra Malaysia (UPM). Hal-hal penting dominan yang terjadi selama pelaksanaan kegiatan ialah keaktifan pelaksana kegiatan dalam menjalankan berbagai aktivitas dalam kegiatan *Visiting Professor* di universitas mitra. Dalam hal ini, Prof. Dr. Badrun Kartowagiran tidak menemui kendala yang berarti ketika menjalankan seluruh rangkaian kegiatan. Faktor yang mempengaruhi minimnya kendala yang ditemui adalah kapabilitas dan kemampuan adaptasi yang baik yang dimiliki oleh Prof. Dr. Badrun Kartowagiran. Semua rangkaian program dalam kegiatan *Visiting Professor* dapat terlaksanan sebagaimana dengan yang telah direncanakan sebelumnya.

B. Tujuan	Maksud dan tujuan kegiatan *Visiting Professor* untuk melaksanakan kerjasama dengan perguruan tinggi di luar negeri dalam bidang pendidikan. Kerjasama dengan Perguruan Tinggi Luar Negeri (PLTN) sebagai salah satu upaya melaksanakan Visi dan Misi UNY, mewujudkan UNY sebagai *World Class University* (WCU), meningkatkan kualitas pendidikan Program Studi PEP di Asia Tenggara dan meningkatkan daya saing internasional.		
C. Sumber Daya (*Inputs*)	**SDM**	**Rencana**	**Realisasi**
	1) Nama Dosen Pengusul: Prof. Dr. Badrun Kartowagiran, M.Pd. **2) Nama Dosen Mitra:** Dr. Siti Salina Mustakim **3) Nama Universitas Mitra dari Luar Negeri:** Universiti Putra Malaysia (UPM)	**1) Nama Dosen Pengusul:** Prof. Dr. Badrun Kartowagiran, M.Pd. **2) Nama Dosen Mitra:** Dr. Siti Salina Mustakim **3) Nama Universitas Mitra dari Luar Negeri:** Universiti Putra Malaysia (UPM)	**1) Nama Dosen Pengusul:** Prof. Dr. Badrun Kartowagiran, M.Pd. **2) Nama Dosen Mitra:** Dr. Siti Salina Mustakim **3) Nama Universitas Mitra dari Luar Negeri:** Universiti Putra Malaysia (UPM) **(Tidak terjadi perubahan)**
	Biaya	**Rencana**	**Realisasi**
	1) Tiket Pesawat multi maskapai Yogyakarta – Malaysia PP (*at coast*) 2) Penginapan untuk 13 malam di Hotel Kampus 3) Lump sam	**1) Tiket Pesawat multi maskapai Yogyakarta – Malaysia PP (*at coast*):** Rp 5.000.000,00 **2) Penginapan untuk 13 malam di Hotel Kampus** Rp 8.500.000,00	**1) Tiket Pesawat multi maskapai Yogyakarta – Malaysia PP (*at coast*):** Rp 5.000.000,00 **2) Penginapan untuk 13 malam di Hotel Kampus** Rp 8.500.000,00

		3) Lump sam Rp 5.200.000,00	**3) Lump sam** Rp 5.200.000,00 **(Tidak terjadi perubahan)**
	Jumlah Biaya	**Rp 18.700.000,00**	**Rp 18.700.000,00** **(Tidak terjadi perubahan)**
	Sumber Dana Dana ini adalah dana yang dikeluarkan oleh PPs UNY dan baru digunakan untuk kegiatan yang ada di Pascasarjana UNY. Sumber dana untuk kegiatan ini ialah DIPA UNY Tahun Anggaran 2019.		
D. Mekanisme dan Rancangan (*Process*)	Kegiatan pelaksanaan kerjasama yang berupa *Visiting Professor,* dilaksanakan pada tanggal 16 sampai 30 Oktober 2019 yang diikuti oleh Prof. Dr. Badrun Kartowagiran selaku Ketua Program Studi Penelitian dan Evaluasi Pendidikan (PEP) Program Pascasarjana (UNY). Pelaksanaan *Visiting Professor* ini mengalami perubahan jika dibandingkan dengan perencanaan awal. Dalam proposal pengajuan, kegiatan ini direncanakan akan dilaksanakan pada tanggal 14 sampai 17 Oktober 2019. Atas pertimbangan pelaksana dan universitas mitra, *Visiting Professor* digeser pelaksanaannya pada tanggal 16 sampai 30 Oktober 2019. Secara rinci, kegiatan ini dapat dilihat pada jadwal yang terlampir dalam laporan ini.		
E. Keluaran (*Output*)	**Uraian**	**Rencana/Target**	**Realisasi**
	Terlaksananya *Visiting Professor* ke Universiti Putra Malaysia (UPM).	Kegiatan terlaksana dengan lancar.	Kegiatan *Visiting Professor* ke Universiti Malaysia (UPM) sudah terlaksana dengan lancar sesuai dengan jadwal dan perencanaan yang sudah disusun sebelumnya.
F. Hasil (*Outcomes*)	**Uraian**	**Rencana/Target**	**Realisasi**
	Meningkatnya kualitas program studi PEP	Program studi PEP menjalin kerjasama dengan PT LN.	Kualitas program studi PEP meningkat, dibuktikan dengan

	terkait dengan kerjasama PT LN.		terjalinnya kerja sama lanjutan dengan UPM, salah satunya dalam hal pelibatan dosen UPM sebagai editor jural yang dikelola PEP.
G. Indikator Keberhasilan	**Uraian**	**Rencana/Target**	**Realisasi**
	Dosen program studi PEP menjadi *Guest Lecturer* di PT LN.	Dosen program studi PEP menjadi *Guest Lecturer* di PT LN.	Dosen program studi PEP menjadi *Guest Lecturer* di PT LN ketika melakukan *Visiting Professor* di UPM.
H. Rencana Tindak Lanjut	Program studi PEP berencana menjadikan kegiatan *Visiting Professor* sebagai kegiatan rutin yang akan dilakukan setiap tahunnya, khususnya *Visiting Professor* yang bekerjasama dengan UPM sebagai PT LN.		

Lampiran

A. Surat Persetujuan *Visiting Professor* dari Direktur PPs UNY

B. Surat Penerimaan *Visiting Professor* dari Universiti Putra Malaysia (UPM)

C. Jadwal Kegiatan *Visiting Professor*

D. Sertifikat Bukti Telah Melakukan *Visiting Professor*

E. Bahan Ajar Perkuliahan

F. Foto-foto Kegiatan

A. Surat Persetujuan *Visiting Professor* dari Direktur Pascasarjana (PPs) Universitas Negeri Yogyakarta

MINISTRY OF RESEARCH, TECHNOLOGY AND HIGHER EDUCATION
UNIVERSITAS NEGERI YOGYAKARTA
GRADUATE SCHOOL
Jalan Colombo No.1., Karangmalang, Yogyakarta, Indonesia 55281
Phone.: 62-274-550835, 550836 Fax. 520326
Web.: pps.uny.ac.id Email: pps@uny.ac.id, humas_pps@uny.ac.id

Reff. No. : 4558/UN34.17/TU/2019 April 5, 2019
Subject : **Invitation for Visiting Professor Program at Universitas Negeri Yogyakarta**

Dean Faculty of Educational Studies
Universiti Putra Malaysia
43400 Seri Kembangan, Selangor, Malaysia

Dear Sir/Madam,

Greeting from Universitas Negeri Yogyakarta!

On behalf of Universitas Negeri Yogyakarta (UNY), I would like to express our warmest greeting while hoping you are always in prosperity, happiness, and good health. UNY has entered into various collaborative arrangements with others to enhance its academic links and cooperation.

One of the collaborative programs being initiated by the Graduate School of UNY is a visiting professor program which provides reciprocal opportunities to receive and send lecturers. Professor Dr. Badrun Kartowagiran and Dr. Siti Salina Binti Mustakim Salina have intensively discussed about this possibility and come into decision to realize exchange visiting professor next academic year of 2019/2020.

In this regards, it is my great pleasure to invite Dr. Siti Salina Binti Mustakim, Senior Lecturer of Faculty of Educational Studies as a visiting professor to UNY for the period of September 9-21, 2019. Her visit will be supported by the Graduate School of UNY and Educational Research and Evaluation Department will be her host during her stay at UNY and will work with her as she engage in the following activities:

1. Being a lecturer and will be partnered with Prof. Dr. Badrun Kartowagiran in delivering the blended-courses of Classroom Evaluation and Educational Program Evaluation for graduate students;
2. Being a reviewer for journal REID and JPEP;
3. Involved in examining master and doctoral candidates;
4. Delivering guest lecture in specific topic for graduate students.

UNY will cover the cost of her travel, accommodation, and subsistence expenses related to her visit. Unfortunately, UNY does not provide insurance services, and this is a responsibility of all visitors to have a comprehensive travel insurance including the medical cover.

We are indeed looking forward to welcoming her to UNY and we do hope she will have a pleasant visit and stay in Yogyakarta Indonesia.

Sincerely yours,

Prof. Dr. Marsigit, M.A.
Director

B. Surat Penerimaan *Visiting Professor* dari Universiti Putra Malaysia (UPM)

FAKULTI PENGAJIAN PENDIDIKAN
FACULTY OF EDUCATIONAL STUDIES

Our Ref: UPM.FPP.500-4/7
Date: 25th April 2019

Professor Dr Badrun Kartowagiran
Head of Educational Research and Evaluation Department
Graduate School
Universitas Negeri Jogjakarta
Jalan Colombo No 1, Karangmalang
Yogyakarta, Indonesia 55281.

Dear Professor Dr. Badrun,

ACCEPTANCE FOR REQUEST AS VISITING PROFESSOR IN THE FACULTY OF EDUCATIONAL STUDIES, UNIVERSITI PUTRA MALAYSIA

First and foremost, thank you for choosing Universiti Putra Malaysia for your attachment placement as a visiting scholar. I am delighted to have you at the Faculty of Educational Studies (FPP), Universiti Putra Malaysia on October 14 – 26, 2019.

Based on your letter, I am happy that your tentative goals include the following;

1. Delivering courses related to classroom evaluation and educational program evaluation, and
2. Discuss collaborative projects with lecturers of the faculty and particularly with Dr. Siti Salina Mustakim. I would also like to propose that you co-author a few papers with our lecturers.

You will be provided a working area in FPP and access to the internet. You will be assisted by Dr. Siti Salina Mustakim during your stay here.

I am made to understand that all costs incurred for your attachment here will be fully covered by your university. I would also like to extent my gratitude and thanks for providing reciprocal opportunities for Dr. Siti Salina to be invited as a visiting lecturer with all expenses paid by Universitas Negeri Jogjakarta on Sept 9 – 21, 2019. Apart from giving guest lectures, she is also expected to be a reviewer of the REID and JPEP journals, and examining master and doctoral candidates.

Thank you again. Please do not hesitate to let me know if you have any further questions. We look forward to collaborating with you!

"WITH KNOWLEDGE WE SERVE"

Yours sincerely,

PROFESSOR DR. AIDA SURAYA MD. YUNUS
Dean, Faculty of Educational Studies
Universiti Putra Malaysia

Cc: Registrar, Universiti Putra Malaysia
Head, Department of Foundation of Education
Dr. Siti Salina Mustakim

Fakulti Pengajian Pendidikan, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor Darul Ehsan, Malaysia
603-8946 8210 603-8948 0119 http://www.educ.upm.edu.my

C. Jadwal Kegiatan *Visiting Professor* Prof. Badrun Kartowagiran di UPM

FAKULTI PENGAJIAN PENDIDIKAN
FACULTY OF EDUCATIONAL STUDIES

SCHEDULE FOR PROFESSOR DR. BADRUN KARTOWAGIRAN
16TH – 30TH OCTOBER 2019
FACULTY OF EDUCATIONAL STUDIES' UNIVERSITI PUTRA MALAYSIA

Time / Venue	Activities	Person in Charge
	Wednesday, 16 October 2019	
KLIA2	Prof. Dr. Badrun Kartowagiran arrived at KLIA2	Dr. Siti Salina Mustakim
	Thursday, 17 October 2019	
Adjustable	Discussion on Teaching Preparation, Consultation, and Curriculum Review	Dr. Siti Salina Mustakim Dr. Norliza Ghazali Dr. Mohd Rozilee Wazir Norjali Wazir
	Friday, 18 October 2019	
International Office Visiting Professor Room	Visit International Office Teaching Preparation	Dr. Siti Salina Mustakim Asc. Prof. Dr. Umi Kalthum Abd. Manaf Prof. Dr. Badrun Kartowagiran
	Monday, 21 October 2019	
12:00 – 14:00 FPP 219A	Lecturing on FCE3501 Learning Assessment Class [Topic: Authentic Assessment]	Dr. Siti Salina Mustakim
	Tuesday, 22 October 2019	
14:00 – 17:00 A1-20	Lecturing on FCE3501 Learning Assessment Class [Topic: How to Construct Item]	Dr. Siti Salina Mustakim
	Wednesday, 23 October 2019	
2.00 – 5.00pm BT225	Lecturing on CNI5003 Assessment in Instruction [Topic: Validity & Reliability]	Dr. Siti Salina Mustakim
	Thursday, 24 October 2019	
Visiting Professor Room	Co-supervision of Postgraduate Students: 1. Candidate: Mr. Tham Jia Hou Thesis Title: The Capability of Moral Education in Enhancing Student's Moral Behavior in Malacca Secondary School 2. Candidate: K. Thilagavathy a/p M Krishnasamy Thesis Title: An Evaluation of the Implementation of IGCSE Curriculum Towards International Assessment Benchmark	Dr. Norliza Ghazali Prof. Dr. Badrun Kartowagiran
	Friday, 25 October 2019	
Visiting Professor Room	Teaching Preparation	Prof. Dr. Badrun Kartowagiran
	Monday, 28 October 2019	
	Public Holiday	
	Tuesday, 29 October 2019	
Faculty of Educational Studies	Discussion on Mutual Collaboration between UPM and UNY	Faculty Dean Dr. Siti Salina Mustakim Prof. Dr. Badrun Kartowagiran Dr. Norliza Ghazali Dr. Mohd Rozilee Wazir Norjali Wazir
	Wednesday, 30 October 2019	
KLIA2	Depart to KLIA2	Dr. Norliza Ghazali Dr. Mohd Rozilee Wazir Norjali Wazir

Fakulti Pengajian Pendidikan, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor Darul Ehsan, Malaysia
(603-8946 8210 603-8948 0119 http://www.educ.upm.edu.my

D. Sertifikat Bukti Telah Melakukan *Visiting Professor*

UPM

UNIVERSITI PUTRA MALAYSIA

BERILMU BERBAKTI

CERTIFICATE OF APPRECIATION

This certificate is awarded to

PROFESSOR DR. BADRUN KARTOWAGIRAN

from Yogyakarta State University, Indonesia for his precious contribution as a Visiting Professor to the development of teaching and learning process in

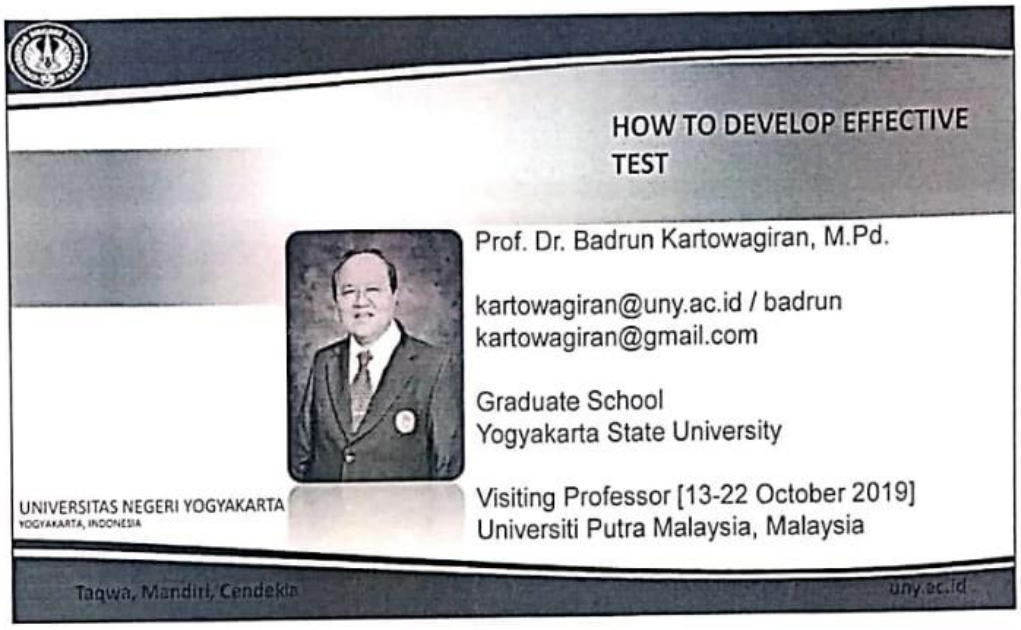
Research and Evaluation

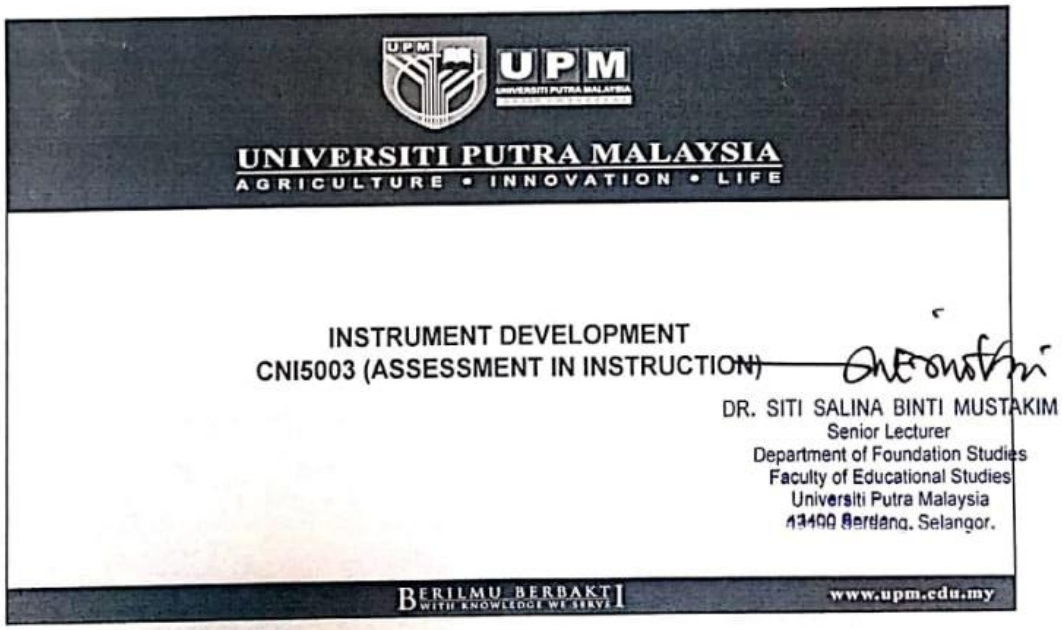
from 16th – 30th October 2019 at Master and Doctoral Programme, Faculty of Educational Studies, Universiti Putra Malaysia

PROFESSOR DR. SAMSILAH ROSLAN
Dean
Faculty of Educational Studies
Universiti Putra Malaysia
43400 Serdang, Selangor
MALAYSIA

E. Bahan Ajar Perkuliahan

1) *How to Develop Effective Test*





HOW TO DEVELOP EFFECTIVE TEST

Prof. Dr. Badrun Kartowagiran, M.Pd.

kartowagiran@uny.ac.id / badrun kartowagiran@gmail.com

Graduate School
Yogyakarta State University

Visiting Professor [13-22 October 2019]
Universiti Putra Malaysia, Malaysia

HOW TO DEVELOP EFFECTIVE TEST

Effective test development requires a systematic, detail-oriented approach based on sound theoretical educational measurement principles.

Effective test development requires a systematic, well-organized approach to ensure sufficient validity evidence to support the proposed inferences from the test scores.

All of details must be well executed to produce a test that estimates examinee achievement or ability fairly and consistently in the content domain purported to be measured by the test and to provide documented evidence in support of the proposed test score inferences.

"There are 12 steps for effective test development according to Steven M. Downing (2011)"

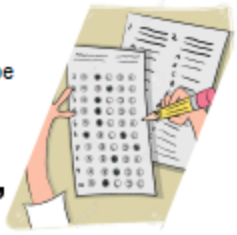
STEP 1. OVERALL PLAN

Every testing program needs some type of overall plan.

The first major decision is: What construct is to be measured? What score interpretations are desired? What test format or combination of formats (selected response or constructed response/performance) is most appropriate for the planned assessment? What test administration modality will be used (paper and pencil or computer based)?

Next

"The key of of Step 1–type tasks and decisions include a clear, concise, well-delineated purpose of the planned test."



Next

The purpose of testing forms an operational definition of the proposed test and guides nearly all other validity-related decisions related to test development activities.

Ultimately, major steps such as content definition, the methods used to define the test content domain, and the construct hypothesized to be measured by the examination are all directly associated with the stated purpose of the test.

Next

"The choice of psychometric model, whether classical measurement theory or item response theory, may relate to the proposed purpose of the test, as well as the proposed use of the test data and the technical sophistication of the test developers and test users."

Next

In many important ways, step 1 is the most important step of the twelve tasks of test development.

A project well begun is often a project well ended.

"Step 1 discuss the importance of:
- clearly defining the purpose of the test
- following careful test development procedures, and
- providing a definitive rationale for the choice of psychometric model for scoring"

Taqwa, Mandiri, Cendekia

uny.ac.id

STEP 2. CONTENT DEFINITION

One of the most important questions to be answered in the earliest stages of test development is: What content is to be tested?

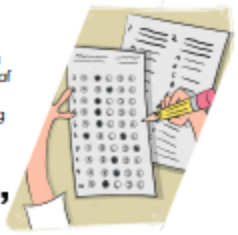
If the content domain is ill defined or not carefully delineated, no amount of care taken with other test development activities can compensate for this inadequacy

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

"Content defining methods vary in rigor, depending on the purpose of the test, the consequences of decisions made from the resulting test scores, and the amount of defensibility required for any decisions resulting from test scores."



Next

Taqwa, Mandiri, Cendekia

uny.ac.id

For some lower stakes achievement tests, the content-defining methods may be very simple and straightforward, such as instructors making informal (but informed) judgments about the appropriate content to test.

For other very high-stakes examination programs, content definition may begin with a multiyear task or job analysis, costing millions of dollars, and requiring the professional services of many testing professionals.

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

For high-stakes achievement examinations, test content defining methods must be systematic, comprehensive, and defensible.

"In this case, strict and defensible methods of defining and depicting domain content are needed; all decisions regarding content, format, and method of content selection become important aspects of proof of validity."

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

The defensibility of the content-defining process is associated with the rigor of the methods employed. One essential feature is the unbiased nature of the methods used to place limits around the universe or domain of knowledge or performance.

"The requirements for defensibility and rigor of content-defining methods is directly proportional to the stakes of the examination and the consequences of decisions made about individuals from the resulting test scores."

Taqwa, Mandiri, Cendekia

uny.ac.id

STEP 3. TE$T SPECIFICATION$: BLUEPRINTING THE TEST

Test specifications refers to a complete operational definition of test characteristics, in every major detail, and thus includes the test blueprint

A test blueprint defines and precisely outlines the number (or proportion) of test questions to be allocated to each major and minor content area and how many (what proportion) of these questions will be designed to assess specific cognitive knowledge levels.

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

Test blueprint = **Table specifications**

Test blueprint used to ensure congruence between classroom instruction and test content
(Wilson, V; Livingston, R; and Reynolds, C., 2009)

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

"At a minimum, the test specifications must describe:
- The type of testing format to be used (selected response or constructed response/performance)
- The total number of test items (or performance prompts) to be created or selected for the test, as well as the type or format of test items
- The cognitive classification system to be used
- Whether or not the test items or performance prompts will contain visual stimuli
- The expected item scoring rules
- How test scores will be interpreted
- The time limit for each item"

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

Example of Test Blueprint

Content Area	*Recall*	*Application*	*Problem Solving*	*Totals*
Content define	4	10	6	20
Test specs	3	8	4	15
Item writing	4	10	6	20
Assembly: print administer	2	5	3	10
Test scoring	3	7	5	15
Test standards	4	10	6	20
Totals	20%	50%	30%	100%

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

STEP 4. ITEM DEVELOPMENT

Early in the test development process, the test developer must decide what test item formats to use for the proposed examination.

The multiple choice format (and its variants), with some ninety years of effective use and an extensive research basis, is the item format of choice for most testing programs (Haladyna, 2004)

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

The multiple-choice item is the workhorse of the testing enterprise, for very good reasons.

The multiple-choice item is an extremely versatile test item form; it can be used to test all levels of the cognitive taxonomy, including very high-level cognitive processes.

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

The multiple-choice item is an extremely efficient format for examinees, but is often a challenge for the item writer

"The choice of item format is a major source of validity evidence for the test. A clear rationale for item format selection is required."

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

Competent content review and professional editing of test questions is also an important source of validity evidence for high-stakes examinations.

- All test items should be written against detailed test specifications
- These draft or raw items should then be reviewed and edited by a professional test editor, who has specialized editorial skills with respect to testing

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

Item Development ➡ Item Writer Training

Training of item writers is an important validity issue for test development.

Without specific training, most novice item writers tend to create poor-quality, flawed, low-cognitive-level test questions that test unimportant or trivial content.

Taqwa, Mandiri, Cendekia

uny.ac.id

STEP 5. TEST DESIGN AND ASSEMBLY

The validity of the final test score interpretation very much relies on the competent and accurate test assembly process.

The errors or serious flaws and omissions in the test assembly process can be obviously glaring and have the potential of seriously reducing the validity evidence for examination scores.

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

For most achievement tests, the most important validity-related issues in test assembly are the correspondence of the content actually tested to the content specifications and the high-level quality control of this entire process.

The test assembly step operationalizes the exacting sampling plan developed in Steps 2 and 3 and lays the solid foundation for inferential arguments relating sample test scores to population or universe scores in the domain.

Taqwa, Mandiri, Cendekia

uny.ac.id

STEP 6. TEST PRODUCTION

All the prior test development work comes to fruition in Step 6

Test production activities and their validity implications apply equally to performance tests and selected-response tests.

Step 6—Test Production—truly operationalizes the examination, making final all test items, their order, and any visual stimuli associated with the test items.

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

All tests, whether large-scale national testing programs or much smaller local testing programs, must ultimately be printed, packaged for computer administration, or published in some form or medium.

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

Security issues are prominent for test production.

Human error is the most likely source of test security breaches, even in this era of high-tech computer-assisted test production.

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

Test production security standards and policies must be developed, implemented, and quality controlled for all high-stakes examinations and should reflect the consequences of testing somewhat proportionally.

Independent audits of these security procedures should be carried out periodically by security professionals, especially for all computer-based security systems.

Taqwa, Mandiri, Cendekia

uny.ac.id

STEP 7. TEST ADMINISTRATION

There are major validity issues associated with test administration because much of the standardization of testing conditions relates to the quality of test administration.

Whether the test is administered in local school settings by teachers, in large multisite venues by professional proctors, or by trained staff at nationwide computer-based testing centers, many of the basic practices of sound test administration are the same

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

The test administration conditions—standard time limits, proctoring to ensure no irregularities, environmental conditions conducive to test taking, and so on—all seek to control extraneous variables in the "experiment" and make conditions uniform and identical for all examinees.

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

Without adequate control of all relevant variables affecting test performance, it would be difficult to interpret examinee test scores uniformly and meaningfully.

This is the essence of the validity issue for test administration considerations.

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

Security is a major concern for test administration.

For most examinations (e.g., large-scale, high-stakes tests), the entire test development process is highly secure, with extremely limited access to test materials.

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

Paper and Pencil Test Administration

For paper-and-pencil test administration, proctoring is one of the most critical issues. For large-scale testing programs, testing agencies typically have a cadre of well-trained and well experienced proctors available to oversee test administration.

Computer Based Test Administration

For large-scale computer-based tests, proctoring is generally delegated to the agency providing the computer test administration network.

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

"The Standards associated with test administration generally deal with issues of standardization and examinee fairness, such as time limits, clarity of directions to examinees, and standard conditions for testing."

Taqwa, Mandiri, Cendekia

uny.ac.id

STEP 8. SCORING EXAMINATION RESPONES

Test scoring is the process of applying a scoring key to examinee responses to the test stimuli.

The responses are not the measurement; rather an application of some scoring rules, algorithms, or rubrics to the responses result in measurement

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

The most obvious issue of scoring relates to accuracy of scoring.

Scoring errors always reduce validity evidence for the test and can invalidate the results.

If final test scores are to have valid meaning, especially the meaning that was anticipated by the test developers (a measure of the construct of interest, an adequate sample of the domain of knowledge, and so on), a scoring key must be applied with perfect accuracy to the examinee item responses.

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

Scoring errors always reduce validity evidence for the test and can invalidate the results.

Validity evidence can be reduced by either a faulty (inaccurate) scoring key or flawed or inaccurate application of the scoring key to responses.

Thus, high levels of quality control of the scoring process are essential to validity.

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

"Scoring can be extremely simple or very complex, depending on the type of test item or stimuli."

The responses to single-best-answer multiple choice items are easily scored by computer software, whereas responses to complex computer simulation problems can be more challenging to score reliably.

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

Scoring Examination Responses

- ❑Preliminary Scoring and Key Validation
- ❑Final Scoring

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

❑Preliminary Scoring

Key validation is the process of preliminary scoring and item analysis of the test data, followed by a careful evaluation of the item-level data to identify potentially flawed or incorrect items prior to final test scoring.

A final "key validation" or key verification step increases the validity evidence for all examinations. This two-step scoring process is essential for tests containing newly written and non-pretested items, because it is possible that such items may contain invalidating flaws that were not detected during the item writing and review process.

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

❑Final Scoring

Final scoring of the examinee responses follows the preliminary scoring and key validation procedures.

The final answer key must be carefully proofread and quality controlled for absolute accuracy.

- ❑ A complete final item analysis includes summary test statistics for the test administration.
- ❑ Summary test statistics are critically important validity evidence and must be thoroughly evaluated and documented.
- ❑ Any anomalies identified by final item analysis or final summary test statistical analyses must be thoroughly investigated and resolved prior to reporting test scores.

Taqwa, Mandiri, Cendekia

uny.ac.id

STEP 9. ESTABLISHING PASSING SCORES

Standard setting is a complex issue with a sound basis in the research literature.

It must be noted that methods and procedures used to establish passing scores may take place at several different stages of test development, depending on the methods used and the overarching philosophy of standard setting adopted by the test developers and users.

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

The basic decisions on the type of standard-setting method to use should be made early in the test development process, because extensive planning may be required to successfully implement certain types of standard setting procedures.

Further, some methods may require multiple exercises or studies, taking place at different times during the test development process and concluding after the test has been administered.

Taqwa, Mandiri, Cendekia

uny.ac.id

STEP 10. REPORTING EXAMINATION RESULTS

Relative standard-setting methods—normative methods—use the actual test performance data, usually from some well-defined group of test takers (e.g., first-time takers of the test) to establish a point on the distribution of test scores to designate as the cut score or passing point.

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

Examinees have a right to an accurate, timely, meaningful, and useful report of their test performance.

Score reports must be written in language that is understandable to recipients and all appropriate cautions and caveats about misuse of test scores must be clearly and unequivocally stated.

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

The score scale used to report test results varies by the type of test, the purpose of the examination, and the sophistication of the examinees.

For score reporting, the choice of raw scores, percent-correct scores, scaled scores, equated scaled scores, or other types of derived scores should be determined solely on the basis of maximizing communication with the examinee.

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

"Whatever score scale is used for reporting, the reported metric should be clearly defined and described in language that is easily understood by the examinee and maximizes the probability of avoiding score misinterpretation."

Taqwa, Mandiri, Cendekia

uny.ac.id

STEP 11. ITEM BANGKING

The process of securely storing test items for potential future use is typically referred to as item banking

Item banks can be as simple as a secure file cabinet using paper copies of examination questions, with the appropriate identifying information and test usage data (such as item difficulty, item discrimination).

Next

Taqwa, Mandiri, Cendekia uny.ac.id

All item banking systems must, at minimum permit the storage, sorting, and retrieval of several variables:

- a unique item identification number
- content classification of the test questions (with several sub classifications of content)
- a cognitive-level classification of the test item
- historical item usage information such as the test form identification (years, dates) of prior use
- the item difficulty and item discrimination indices for each prior use (Item Response Theory parameters, if appropriate)

Taqwa, Mandiri, Cendekia uny.ac.id

STEP 12. TE$T TECHNICAL REPORT

The technical report is the culminating test development activity and serves the major, but often ignored, purpose of providing thorough documentation of all the validity evidence for a test, identifies potential threats to validity, and makes recommendations for improvement in the testing program that may strengthen validity evidence.

Next

Taqwa, Mandiri, Cendekia uny.ac.id

The overall quality of tests can be improved by focusing careful attention on technical reporting.

The examination technical report is also useful in independent evaluations of testing programs, providing a convenient and systematic summary of all important test development activities for review

Next

Taqwa, Mandiri, Cendekia uny.ac.id

"Technical reports must be developed such that all important validity evidence for the testing program is systematically documented in a manner that is easily accessible to all who have a legitimate need to access this information."

Taqwa, Mandiri, Cendekia uny.ac.id

REFERENCE

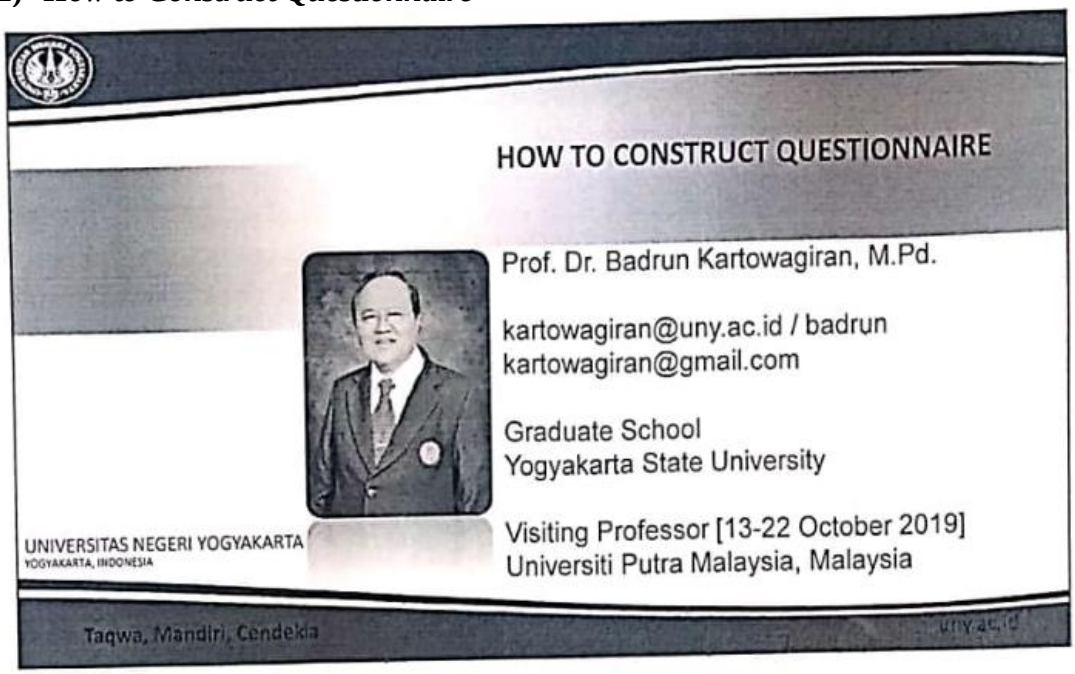
"Downing, S.M., & Haladyna, T.M. (2011). *Handbook of Test Development*. Lawrence Erlbaum Associates, Publisher"

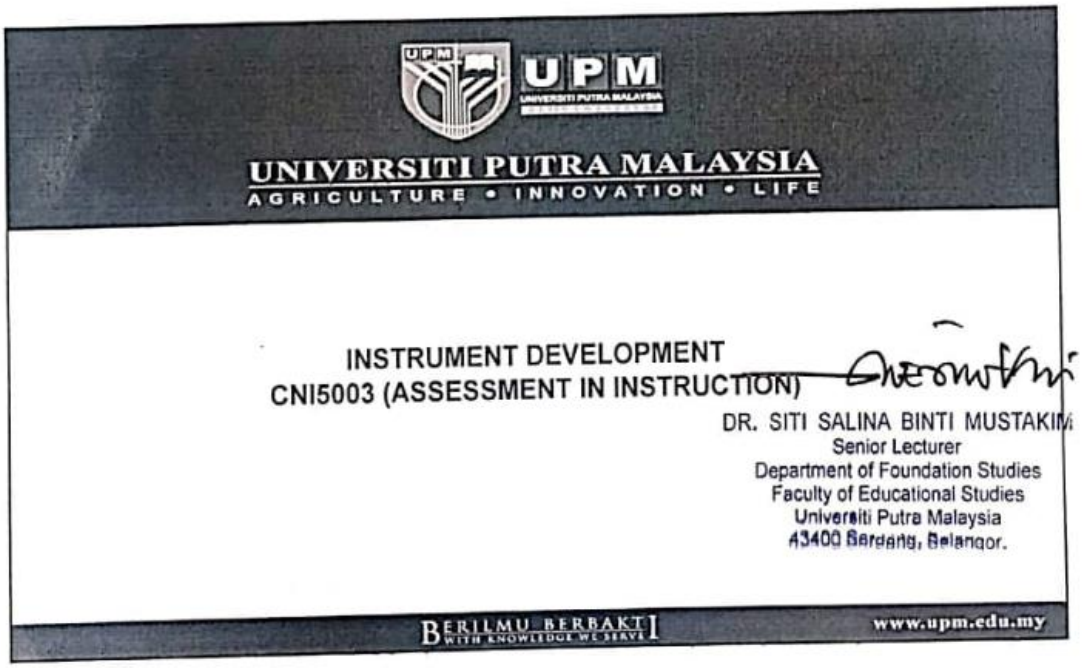
Taqwa, Mandiri, Cendekia uny.ac.id

THANK YOU

Taqwa, Mandiri, Cendekia uny.ac.id

2) ***How to Construct Questionnaire***





HOW TO CONSTRUCT QUESTIONNAIRE

Prof. Dr. Badrun Kartowagiran, M.Pd.

kartowagiran@uny.ac.id / badrun kartowagiran@gmail.com

Graduate School
Yogyakarta State University

Visiting Professor [13-22 October 2019]
Universiti Putra Malaysia, Malaysia

HOW TO CONSTRUCT QUESTIONNAIRE

The goal of the questionnaire is to tap into and understand the opinions of the research participants about variables related to your research objectives

As you construct your questionnaire, you must ask yourself if your questions will provide clear data about what your participants think or feel

"There are 15 principles in compiling a questionnaire according to Johnson Burke (2019)

Principle 1. Make Sure The Questionnaire Items Match Your Research Objectives

You must always determine why you intend to conduct your research study before you can write questionnaire

If you plan to conduct an explanatory research study, your questionnaire will usually not need to be as detailed and specific as if you plan to conduct confirmatory research

Next

When your primary goal is explore the topic, you want to be broad in your questions that you do not miss an important concept that your research participants feel is relevant

In both exploratory and confirmatory research, you should carefully review the existing research literature, as will as any related instruments that have already been used for your research objectives, before deciding to construct your own questionnaire

Next

Important Note

One of the worst things that can happen in questionnaire-based research is to realize that you should have asked a question or included a variable **after** your data have bee collected

This omission of a question was not asked about an important issue could indicate that the design of the questionnaire did not carefully consider the research on the topic before designing the questionnaire

As a result, a likely important variable was not measured fully, which will affect the research result as well as the researchers' understanding of the topic

Principle 2. Understand Your Research Participant

A key to effective questionnaire construction is understanding your research participants

Remember that is the participants, not you, who will be filling out the questionnaire

A very important strategy when you write questionnaire is to develop an emphatic understanding or an ability to "think like your potential research participants"

If the questionnaire does not "make sense" to your participants, it will not work

Principle 3. Use Natural and Familiar Language

You should use language that is understandable to be people who are going to fill out your questionnaire. You must know enough about your participants to use language familiar to them. Consider the age of your participants, their educational level, and any of their relevant cultural characteristics when deciding on the kind of language to use.

Remember!
That is very possible that not everyone uses the same everyday language as you. So, that not every participants of the research uses same languagage as the researcher

Next

✓The use of natural and familiar language makes it easier for participants to fill iota questionnaire and helps participants feel more relaxed and less threatened by the task of filling it out

✓One key issue related to both the principle of understanding your participants and that of using natural and familiar language is determining an appropriate reading level

Next

"When construct a questionnaire, important to use the reading level that is natural and appropriate for your research participants. Poorly constructed questionnaire are written at either too high or too low a reading level for the intended participants.

Reading Level is Important

Principle 4. Write Items that Are Clear, Precise, and Relatively Short

Each item on your questionnaire should be understandable to you (the researcher) and to the participants (the people filling out the questionnaire).

If the participants are not clear about what is being asked of them, their responses will result in data that cannot or should not be used in a research study

Write Items that Are Clear, Precise, and Relatively Short

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

Your goal is for each research participant to interpret the meaning of each item in the questionnaire in exactly the same ways

If you must use technical term, remember to define it for the participants.

Try to keep most items relatively short because long item can be confusing and stressful for research participants

Write Items that Are Clear, Precise, and Relatively Short

Taqwa, Mandiri, Cendekia

uny.ac.id

Principle 5. Do not Use "Leading" or "Loaded" Questions

A leading or loaded questions biases the response the participant gives to the question.

A loading question is one that is phrased in such a way it suggest a certain answer.

For example, the emotionally charged word liberal was often avoided by politicians with left-of-center leanings during the 1980s and 1990s because the word created a negative reaction in some people regardless of the content of the statement.

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

Here is an example of a loading question:

Don't you agree that teachers should earn more money than they currently earn?
- ☐ Yes, they should earn more
- ☐ No, they should not earn more
- ☐ Don't know/no opinion

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

The phrase "Don't you agree" leads the participant. A more neutral wording of this question would be as follows:

Do you believe teacher salaries are lower than they should be, higher than they should be, or at the right amount?
- ☐ Teacher salaries are lower than they should be
- ☐ Teacher salaries are higher than they should be
- ☐ Teacher salaries at the right amount
- ☐ Don't know/no opinion

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

Here is an entertaining example of a question that is leading and has loaded phrases in it (from Bonevac, 1999)'

Do you believe that you should keep more money of your hard-earned or that government should get more of your money for increasing bureaucratic government programs?
- ☐ Keep more of my hard-earned money
- ☐ Give my money to increase bureaucratic government programs
- ☐ Don't know/no opinion

"Always remember that your goal is to write questionnaire items that help participants feel free to provide their natural and honest answers. You want to obtain responses that are undistorted by the wording of the questions"

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

Principle 6. Avoid Double-Barreled Questions

A double-barreled question combines two or more issues or attitude objects in a single item.

Here's an example:

Do you think that teachers should have more contact with parents and school administrators?

A you can see, this single item asks about two different issues. The questions is really asking, *Do you think that teachers should have more contact with parents?* And *Do you think that teachers should have more contact with administrators?*

Taqwa, Mandiri, Cendekia

uny.ac.id

Principle 7. Avoid Double Negatives

Double negative: A sentence construction that includes two negatives

When participants are asked for their agreement statement, double negatives can easily occur.

For example:

Do you agree or disagree with the following statement?

Teachers should not be required to supervise their students during library time.

If you disagree the statement, you must construct a double negative (a sentence construction that includes two negatives). If you disagree, you are saying that you do not think that teachers should not supervise students during library time (Converse & Presser, 1986). In the other words, you probably believe that teachers should supervise during library time.

Taqwa, Mandiri, Cendekia

uny.ac.id

Principle 8. Determine Whether an Open-Ended or a Closed-Ended Is Needed

Quantitative questionnaire: A questionnaire based on closed-ended items and typically used in confirmatory or quantitative research

Qualitative questionnaire: A questionnaire based on open-ended items and typically used in exploratory or qualitative research

Mixed questionnaire: A questionnaire that includes a mixture of open-ended and closed-ended items

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

Open-ended questions that allows participants to respond to their own words

Closed-ended questions: a questions that forces participants to choose from a set of predetermined responses

Item stem: The set of words forming a question or statement

Taqwa, Mandiri, Cendekia

uny.ac.id

Principle 9. Use Mutually Exclusive and Exhaustive Response Categories for Closed-Ended Questions

Categories are mutually exclusive when they do not overlap. For example, that following response categories for a question about the participant's age are not mutually exclusive:

- ☐10 or less
- ☐10 to 20
- ☐20 to 30
- ☐30 to 40
- ☐40 to 50
- ☐50 to 60
- ☐60 to 70
- ☐70 to 80
- ☐80 or greater

Next

Taqwa, Mandiri, Cemerlang — uny.ac.id

Do you see the problem with the response categories?

The problem is that they overlap

For example, a person who is 20 years old could be placed into two categories. In fact, persons aged 10, 20, 30, 40, 50, 60, 70, and 80 can all be placed into more than one category. In short, the response categories are not mutually exclusive. In a moment, we will show you how to fix this problem.

Next

Taqwa, Mandiri, Cemerlang — uny.ac.id

A set response categories is exhaustive when there is a category available for all legitimate responses. For example, what is the problem with the following categories from a question asking for your current age?

- ☐1 to 4
- ☐5 to 9
- ☐10 to 14

The problem is that these three categories are not exhaustive because there is no category available for anyone over the age of 14 or anyone younger 1 year old. A set of categories is not exhaustive unless there is a category available for all potential responses.

Next

Taqwa, Mandiri, Cemerlang — uny.ac.id

Putting the ideas of mutually exclusive categories together, you can see that the following set of response categories is mutually exclusive and exhaustive:

Which of the following categories includes your current age?

- ☐Less than 18
- ☐18 to 29
- ☐30 to 39
- ☐40 to 49
- ☐50 to 59
- ☐60 to 69
- ☐70 to 79
- ☐80 to older

The principle of mutually exclusive categories applies because none of the categories is available for every possible age. Whenever you write a standard closed-ended question (a question with an item stem and a set of predetermined response categories), remember to make sure that your response categories are mutually exclusive and exhaustive

Taqwa, Mandiri, Cemerlang — uny.ac.id

Principle 10. Consider The Different Types of Response Categories Available for Closed-Ended Questionnaire Items

In this section, we introduce several popular types of closed-ended response categories by explaining the ideas of rating scale, rankings, semantic differentials, and checklist

Next

Taqwa, Mandiri, Cemerlang — uny.ac.id

- ✓**Rating Scale:** A continuum of response choices
- ✓**Numerical rating scale:** A rating scale that includes a set of numbers with anchored endpoints
- ✓**Anchor:** A written descriptor for a point on a rating scale
- ✓**Fully anchored rating scale:** A rating scale on which all points are anchored

Next

Taqwa, Mandiri, Cemerlang — uny.ac.id

Rating Scales

Rating scales produce numerical (quantitative) data rather than qualitative data (nominal-level data)

Rating scales have been used by researchers for quite a long time. In an early review of the history of rating scales, Guilford (1936) provided examples from as early as 1805 and many other examples from shortly after 1900. Some important early developers of rating scales were Sir Francis Galton (1822-1911), Karl Pearson (1857-1936), and Rensis Likert (1903-1981).

Next

Taqwa, Mandiri, Cemerlang — uny.ac.id

A numerical rating scales

A numerical rating scales consists of a set of numbers and "anchored" endpoints. When you anchor a point on a rating scale, you label the point with a written descriptor. Here is an example of an item stem and a numerical rating scale with anchored endpoints:

How would you rate the overall job performance of your school principal?

1	2	3	4	5	6	7
Very low						Very high

Next

Taqwa, Mandiri, Cemerlang — uny.ac.id

As you can see, the first endpoint (1) is enchored with the words very low.

The other endpoint (7) is enchored with the words very high.

This is a 7 point rating scale because there is a total of seven points on the scale

If you see an even number of points, a respondent might misinterpret one of the two centremost numbers as representing the center of neutral point (Dillman, 2007).

If you choose to use an even number of points, you will need to anchor the two centremost numbers or clearly anchor the area between the two centremost numbers.

Next

Taqwa, Mandiri, Cemerlang — uny.ac.id

A similar type of rating scale is called a fully anchored rating scale. A fully anchored rating scale has all points anchored with descriptors.

1	2	3	4	5
Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree

This scale called a 5-point rating scale because there are five points on the scale. Regardless, you should attempt to make the words or anchor used for adjacent points an equal distance apart from each other. You must be very careful in your choice of anchors for both fully and partially anchored scales.

Taqwa, Mandiri, Cemerlang — uny.ac.id

Principle 11. Use multiple items to measure abstract constructs

Multiple items designed to measure a single construct are used to increase the reliability and validity of the measure. Perhaps the most commonly used procedure for the measurement of abstract constructs is a summated rating scale (also called a Likert scale).

Summated rating scale A multi-item scale that has the responses for each person summed into a single score

Likert scale A type of summated rating scale invented by Rensis Likert

Next

Taqwa, Mandiri, Cendekia uny.ac.id

"The key advantages of multiple-item rating scales compared to single-item rating scales are that multiple-item scales provide more reliable (i.e., more consistent or stable) scores and they produce more variability, which helps the researcher make finer distinctions among the respondents

Next

Taqwa, Mandiri, Cendekia uny.ac.id

If you want to measure a complex construct (such as self-efficacy, locus of control, risk taking, test anxiety, dogmatism, or temperament), the use of a multiple-item scale is pretty much a necessity. When you want to measure constructs such as these, you should not, however, jump to develop your own scale. Rather, you should conduct a literature search to find already validated measures of your construct.

If a measure is not available, only then would you need to consider developing your own measure. The development of a good summated rating scale takes a lot of time and expertise, and extensive validation is required before the scale should be used in a research study

Next

Taqwa, Mandiri, Cendekia uny.ac.id

Principle 12. Consider Using multiple methods when measuring abstract construct

Have you found that there is one type of measurement on which you do better on than others? For instance, do you usually do well on essay tests, no matter the topic, but do worse on true/false tests?

If you have experienced something like this, you have seen why Principle 12 is important

Next

Taqwa, Mandiri, Cendekia uny.ac.id

The use of multiple measurement methods is so important today that more and more researchers are using "measurement models" based on two or even three measurement methods or procedures (e.g., questionnaires, interviews, observations, standardized tests).

The point is that the more methods a researcher uses to measure the relevant concepts or constructs, the more confidence you can place in the researcher's ability to tap into the characteristics of the concept, rather than the method.

Taqwa, Mandiri, Cendekia uny.ac.id

Principle 13. Use caution if you reverse the wording in some of the items to prevent response sets in multi-item scales

When participants rate multiple items using the same or similar rating scale, a "response set" might occur.

A response set is the tendency for a research participant to respond to a series of items in a specific direction, regardless of the differences in item content.

Next

Taqwa, Mandiri, Cendekia uny.ac.id

One type of response set is called the acquiescence response set, which is the tendency to say yes rather than no or to agree rather than to disagree on a whole series of items.

Another response set, called the social desirability response set, is the tendency to provide answers that are socially desirable.

Next

Taqwa, Mandiri, Cendekia uny.ac.id

"One technique used to help prevent response sets (especially the acquiescence response set) is to reverse the wording (and scoring) in some of the items

Taqwa, Mandiri, Cendekia uny.ac.id

Principle 14. Develop a questionnaire that is properly organized and easy for the participant to use

When constructing a questionnaire, you should begin the questionnaire with positive or nonthreatening items because doing so helps obtain commitment from participants as they fill out the questionnaire.

Demographic questions should generally go last in a questionnaire, with a lead-in such as "To finish this questionnaire, we have a few questions about you."

Next

Taqwa, Mandiri, Cendekia uny.ac.id

You should include clear instructions throughout your questionnaire and not put too many items on a page.

If a questionnaire has several topical sections, you should provide transitional or "lead-in" statements to orient the participants to each new topic.

Other important tips are to give your questionnaire a title, number the items consecutively from the beginning to the end, list response categories vertically rather than horizontally (rating scales can be done horizontally or vertically), provide an open-ended question at the end of your questionnaire

Next

Taqwa, Mandiri, Cendekia uny.ac.id

Always try to make your questionnaire look professional, because participants are more likely to fill it out and they will go away with a better impression of you and your organization.

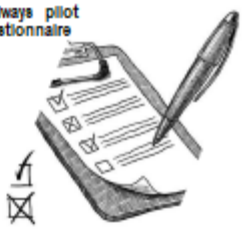
"Remember that the appearance and quality of your questionnaire also reflect on you and your organization."

Taqwa, Mandiri, Cendekia

uny.ac.id

Principle 15. Always pilot test your questionnaire

"It is a cardinal rule in research that you must "try out," or pilot test, your questionnaire to determine whether it operates properly before using it in a research study."



Next

Taqwa, Mandiri, Cendekia

uny.ac.id

You should conduct your pilot test with a minimum of 5 to 10 people. You may want to start with colleagues or friends, asking them to fill out the questionnaire and note any points of confusion.

"You will need to pilot test the questionnaire with several individuals similar to those who will be in your research study."

Taqwa, Mandiri, Cendekia

uny.ac.id

The Steps in Constructing Questionnaire

STEP 1

Review the relevant literature and begin planning the questionnaire

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

Remember that if a questionnaire that fits your needs is already available, than there is no need to construct a new questionnaire

"Think about:
- ❑ Do you understands the targeted participants?
- ❑ Do you understands the issue to be examined?
- ❑ What variables do you wants to measure?
- ❑ What do you want to know in the participans' own word?"

Taqwa, Mandiri, Cendekia

uny.ac.id

STEP 2

Write the items for the questionnaire

Think about:
- ❑ Have you examined other related questionnaire?
- ❑ Have you examined items on other high-quality questionnaires that will helpful models as you write yours?
- ❑ Have you asked others (friend, family members, students) if your items are clear?

Taqwa, Mandiri, Cendekia

uny.ac.id

STEP 3

Design the layout and overall the questionnaire

Think about:
- ❑ Does the questionnaire have a title, clear directions, sections lead-ins, proper section ordering, demographic at the end, and a "thank you" at the conclusion?
- ❑ Have you asked others (collegues) to critique your questionnaire?

Taqwa, Mandiri, Cendekia

uny.ac.id

STEP 4

Conduct a pilot test of the questionnaire

Think about:
- ❑ What people can you administer the questionnaire to who are similar to the kinds of people to be used in your research study?
- ❑ Have I collected validity and reliability data?

Taqwa, Mandiri, Cendekia

uny.ac.id

STEP 5

Administer your questionnaire in your research study

Think about:
- ❑ Does the questionnaire work properly with your research participants?
- ❑ How good are the reliability and validity data with the real participants?
- ❑ Do any items need improvement?

Taqwa, Mandiri, Cendekia

uny.ac.id

Putting It All Together

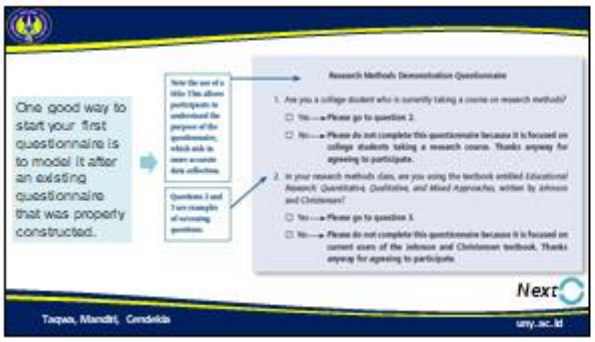
You now have 15 principles of questionnaire construction and steps to build it

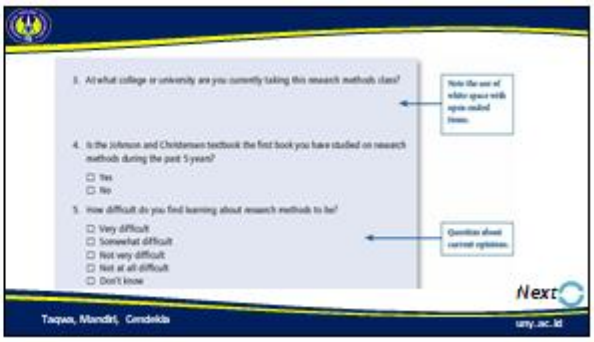
You should feel ready to start the construction of your own questionnaire!

Next

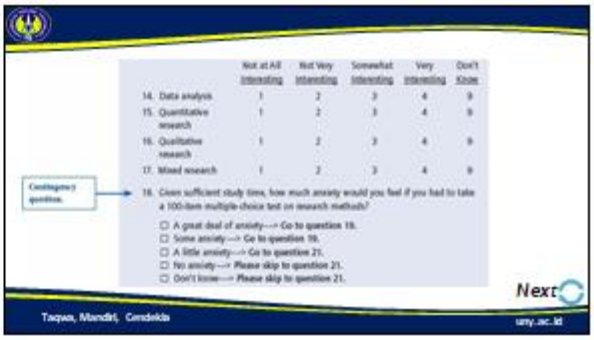
Taqwa, Mandiri, Cendekia

uny.ac.id

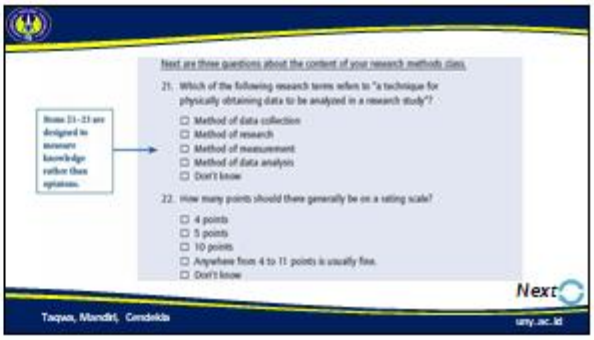

One good way to start your first questionnaire is to model it after an existing questionnaire that was properly constructed.
Research Methods Demonstration Questionnaire
Next
Taqwa, Mandiri, Cendekia
uny.ac.id

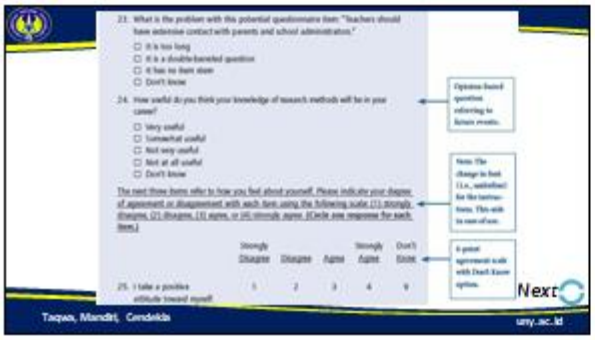

Next
Taqwa, Mandiri, Cendekia
uny.ac.id

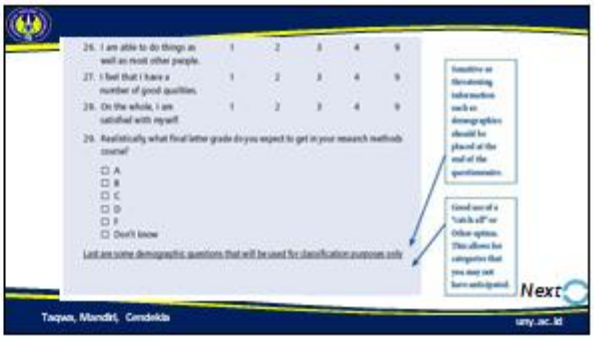

QUESTIONNAIRE
Very often
Sometimes
Always
Next
Taqwa, Mandiri, Cendekia
uny.ac.id

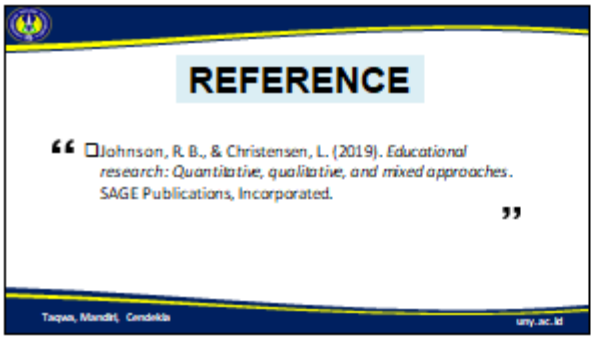

Next
Taqwa, Mandiri, Cendekia
uny.ac.id


Open-ended exploratory question.
19. What do you think are some reasons for your test anxiety?
20. What might be done by your teacher to help reduce your test anxiety?
Next
Taqwa, Mandiri, Cendekia
uny.ac.id


Next
Taqwa, Mandiri, Cendekia
uny.ac.id

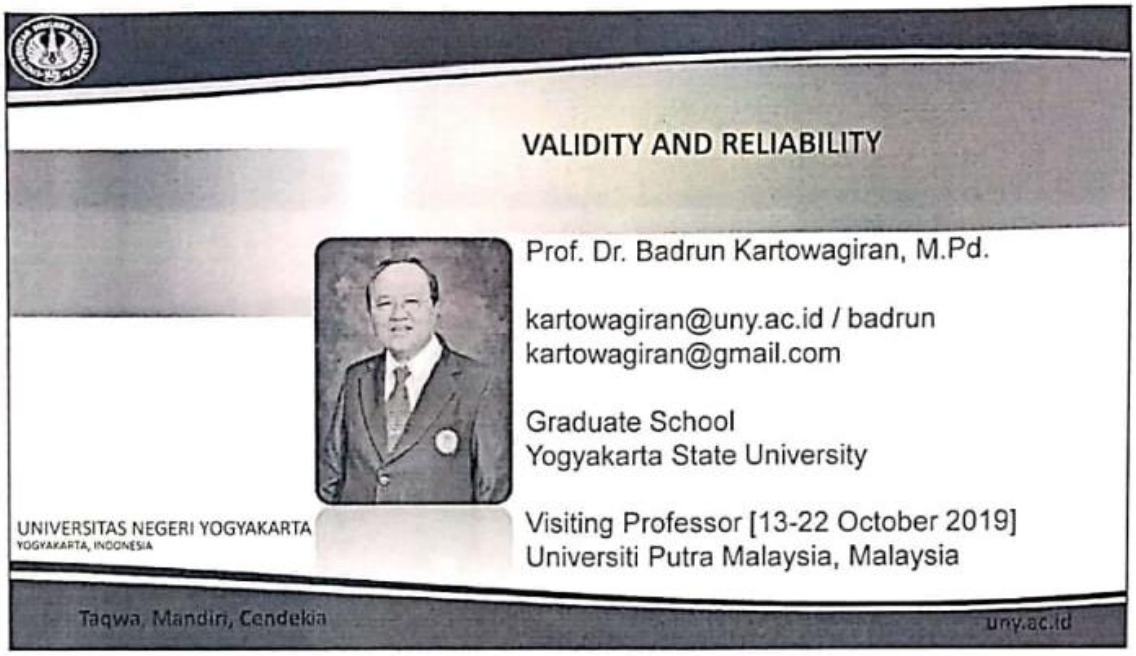

Next
Taqwa, Mandiri, Cendekia
uny.ac.id


Next
Taqwa, Mandiri, Cendekia
uny.ac.id


REFERENCE
Johnson, R. B., & Christensen, L. (2019). Educational research: Quantitative, qualitative, and mixed approaches. SAGE Publications, Incorporated.
Taqwa, Mandiri, Cendekia
uny.ac.id


THANK YOU
Taqwa, Mandiri, Cendekia
uny.ac.id

3) ***Validity and Reliability (in General)***



VALIDITY AND RELIABILITY

Prof. Dr. Badrun Kartowagiran, M.Pd.

kartowagiran@uny.ac.id / badrun kartowagiran@gmail.com

Graduate School
Yogyakarta State University

Visiting Professor [13-22 October 2019]
Universiti Putra Malaysia, Malaysia

Taqwa, Mandiri, Cendekia

uny.ac.id

Validity

"**Validity** is a characteristic that refers to the appropriateness of the inferences, uses, and consequences that result from the assessment.

"**Research Validity** is The correctness or truthfulness of an inference that is made from the results of a study (Johnson Burke, 2019)

Taqwa, Mandiri, Cendekia

uny.ac.id

Taqwa, Mandiri, Cendekia

uny.ac.id

Internal Validity

"**Internal (or causal) validity**

The ability to infer that a causal relationship exists between two variables

Taqwa, Mandiri, Cendekia

uny.ac.id

Internal Validity

Internal (or causal) validity

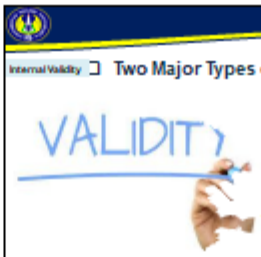
The ability to infer that a causal relationship exists between two variables

Although research is conducted for the multiple purposes of description, exploration, explanation, prediction, and influence, a large amount of research focuses on the goal of attempting to determine whether a causal relationship exists between the independent and dependent variables being investigated

Taqwa, Mandiri, Cendekia

uny.ac.id

Internal Validity ❑ Two Major Types of Causal Relationships



Shadish et al. (2002) have pointed out that there are two types of causal relationships: causal description and causal explanation

Causal Description:

Causal description refers to describing the consequences of manipulating an independent variable

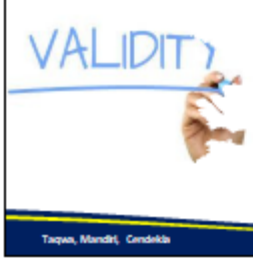
Causal Explanation:

Causal description refers to describing the consequences of manipulating an independent variable

Taqwa, Mandiri, Cendekia

uny.ac.id

Internal Validity ❑ Criteria for Inferring Causation



Three types of evidence are needed to reach a conclusion of causation (i.e., that changes in your independent variable produce changes in your dependent variable).

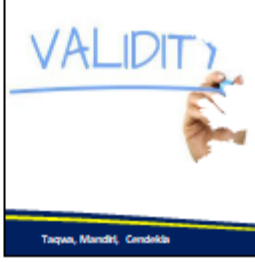
Condition 1:

Condition 1 also called the relationship condition. First, you need evidence that the independent and dependent variables are associated or correlated or related. Do changes in the independent variable correspond to changes in the dependent variable?

Next

Taqwa, Mandiri, Cendekia

uny.ac.id



Condition 2:

Condition 2 also called the temporal antecedence condition. The second type of evidence needed to infer causation is the correct temporal ordering of the variables being investigated, because a cause must precede an effect

Condition 3:

Condition 3 also called the lack of alternative explanation condition. The third type of evidence needed is that the variables being investigated are the ones that are causally related rather than being caused by some confounding extraneous variable.

Taqwa, Mandiri, Cendekia

uny.ac.id

Internal Validity ❑ Threats to Internal Validity in Single-Group Designs

To infer that one variable caused an effect observed in another variable, we must control for all other possible causes.

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

These other possible causes are threats to internal validity because they represent rival or competing or alternative explanations for the results obtained. When such alternative explanations exist, it is impossible to reach a causal explanation with any degree of certainty, leading to highly suspect results that cannot and should not be taken seriously.

"This is why it is necessary to control for and eliminate the systematic influence of these threats."

Taqwa, Mandiri, Cendekia

uny.ac.id

Internal Validity

Threats to Internal Validity in Multigroup Designs

Adding a control group to a single-group design produces a multigroup research design.

The addition of a control group (i.e., moving from a one-group design to a multigroup design) enables you to untangle the confounding effect of the basic threats from the influence of the independent variable. As long as the effect of a basic threat (e.g., history, maturation, testing, instrumentation, or regression artifact) occurs for both groups, it will not cause a problem in the multigroup design because you are determining the treatment effect by comparing the treatment group with a control group.

Taqwa, Mandiri, Cendekia

uny.ac.id

Eksternal Validity

"External validity is a term coined by Campbell and Stanley (1963) and extended by Shadish et al. (2002) to refer to the extent to which the results of a study can be generalized to and across populations of persons, settings, times, outcomes, and treatment variations.

Taqwa, Mandiri, Cendekia

uny.ac.id

External Validity

Population Validity

Population validity refers to the ability to generalize from the sample of individuals on which a study was conducted to the larger target population of individuals and across different subpopulations within the larger target population.

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

The target population is the larger population, such as all children with a learning disability, to whom the research study results are to be generalized. Within this larger target population, there are many subpopulations, such as male and female children with a learning disability.

"Population validity, therefore, has the two components of generalizing from a sample to a target population and generalizing from a sample across the types of persons in the target population."

Taqwa, Mandiri, Cendekia

uny.ac.id

External Validity

Ecological Validity

Ecological validity refers to the ability to generalize the results of a study across settings.

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

"For example:

One study might be conducted in a school whose computers are slow and antiquated. If the results obtained from this study can be generalized to other settings, such as a school well equipped with state-of-the-art technology, then the study possesses ecological validity. Ecological validity therefore exists to the extent that the study results are independent of the setting in which the study was conducted.

Taqwa, Mandiri, Cendekia

uny.ac.id

External Validity

Temporal Validity

Temporal validity refers to the extent to which the results of a study can be generalized across time.

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

"For example:

Thorkildsen, Nolen, and Fournier (1994) assessed children's views of several practices that teachers use to influence motivation to learn.

The data for this study were collected by interviewing 7- to 12-year-old children at one point in time. Although the data are valid for the time period in which they were collected, there is no assurance that the same results would hold true 15 years later.

Taqwa, Mandiri, Cendekia

uny.ac.id

External Validity

Treatment Variation Validity

Treatment variation validity refers to the ability to generalize the results across variations of the treatment. Treatment variation validity is an issue because the administration of a treatment can vary from one time to the next.

Taqwa, Mandiri, Cendekia

uny.ac.id

Outcome Validity

Treatment Variation Validity

Outcome validity refers to the ability to generalize the results across different but related dependent variables. Outcome validity refers to the extent to which the independent variable influences a number of related outcome measures.

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

"For example:

A job-training program is expected to increase the likelihood of getting a job after graduation. This is probably the primary outcome measure of interest. However, an equally important issue is maintaining the job. This means that the person must arrive on time, not miss work, and follow orders as well as demonstrate an acceptable level of performance.

Next

Taqwa, Mandiri, Cendekia uny.ac.id

Sometimes one outcome measure demonstrates that the treatment was effective. However, other outcome measures show no effect and maybe even a negative effect. Using several outcome measures is always desirable because this gives a more complete picture of the overall effect of the treatment.

"Fortunately, this is one of the easier design features to implement. You just need to include several related dependent variables in your study to answer questions about generalizability across outcomes."

Taqwa, Mandiri, Cendekia uny.ac.id

Construct Validity

"The extent to which a higher-order construct is accurately represented in a particular study

Taqwa, Mandiri, Cendekia uny.ac.id

So how do we achieve construct validity?

Construct validity is fostered by having a good definition and explanation of the meaning of the construct of interest. However, every construct, such as violence, has multiple features, and this creates difficulty in identifying the prototypical features of a construct.

Taqwa, Mandiri, Cendekia uny.ac.id

Reliability

"**Reliability/precision** is concerned with the extent to which the scores are free from error (noise). **(McMillan, 2018)**

Reliability refers to the consistency with which a test measures whatever it's measuring. **(Popham, 2017)**

"**Reliability**
The consistency or stability of test scores. **(Johnson Burke, 2014)**

Taqwa, Mandiri, Cendekia uny.ac.id

Reliability is often calculated by using some type of correlation coefficient. When we calculate a correlation coefficient as our measure of reliability, we call it a reliability coefficient.

Taqwa, Mandiri, Cendekia uny.ac.id

A reliability coefficient of zero stands for no reliability at all. (If you get a negative correlation, treat it as meaning no reliability and that your test is faulty.) A reliability coefficient of +1.00 stands for perfect reliability.

"Researchers want reliability coefficients to be strong and positive (i.e., as close to 1.00 as possible) because this indicates high reliability."

Taqwa, Mandiri, Cendekia uny.ac.id

☐ Test-Retest Reliability

Test-retest reliability
A measure of the consistency of scores over time

For example, if you were to assess the reliability of the scores from an intelligence test using the test-retest method, you would give the test to a group of, say, 100 individuals on one occasion, wait a period of time, and then give the same intelligence test to the same 100 individuals again.

Next

Taqwa, Mandiri, Cendekia uny.ac.id

Then you would correlate the scores on the first testing occasion with the scores on the second testing occasion. If the individuals who received high IQ scores on the first testing occasion received high IQ scores on the second testing occasion and the individuals who received low IQ scores on the first testing occasion also received low IQ scores on the second testing occasion, the correlation between the scores on the two testing occasions would be high, indicating that the test scores were reliable.

"If these individuals received very different scores on the two testing occasions, the correlation between the two sets of scores would be low, indicating that the test scores were unreliable."

Taqwa, Mandiri, Cendekia uny.ac.id

☐ Equivalent-Forms Reliability

Equivalent-forms reliability
The consistency of a group of individuals' scores on alternative forms of a test measuring the same thing

Next

Taqwa, Mandiri, Cendekia uny.ac.id

Have you ever taken an exam in which some people got one form of the test and other people got a different form of the test?

If so, you have experienced the use of alternative forms.

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

Equivalent-forms reliability refers to the consistency of a group of individuals' scores on alternative forms of a test designed to measure the same characteristic. Two or more versions of a test are constructed so that they are identical in every way except for the specific items asked on the tests.

Once the two equivalent tests have been constructed, they are administered concurrently to a group of individuals, or the second test is administered shortly after the first test. Either way, each person takes both tests and has scores on both tests. The two sets of scores (participants' scores on each form) are then correlated. This correlation coefficient shows the consistency of the test scores obtained from the two forms of the test.

Taqwa, Mandiri, Cendekia

uny.ac.id

- Internal Consistency Reliability

Internal consistency
The consistency with which the items on a test measure a single construct

Internal Consistency Reliability:
- Split Half
- Coefficient Alpha

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

Internal consistency measures are convenient and are very popular with researchers because they only require a group of individuals to take the test one time.

Next

You do not have to wait for a period of time to elapse after administering the test before you can give it again (as in test-retest reliability), and you do not have to construct two equivalent forms of a test (as in equivalent-forms reliability).

Taqwa, Mandiri, Cendekia

uny.ac.id

- Split-Half Reliability

Split-half reliability
A measure of the consistency of the scores obtained from two equivalent halves of the same test

Once you have created the two halves, reliability of the scores is determined using the following steps:
1. Score each half of the test for every person to whom it was administered.
2. Compute the correlation between scores on the two halves of the test.
3. Adjust the computed correlation coefficient using the Spearman-Brown formula

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

- Coefficient Alpha

Coefficient alpha
A formula that provides an estimate of the reliability of a homogeneous test or an estimate of the reliability of each dimension in a multidimensional test

Coefficient alpha (also called Cronbach's alpha) provides a reliability estimate that can be thought of as the average of all possible split-half correlations, corrected by the Spearman-Brown formula.

Taqwa, Mandiri, Cendekia

uny.ac.id

- Interscorer Reliability

Interscorer reliability
The degree of agreement or consistency between two or more scorers, judges, or raters

The simplest way to determine the degree of consistency between two raters in the scoring of a test or some other performance measure is to have each rater independently rate the completed tests and then compute the correlation between the two raters' scores.

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

For example, assume that you had each student in a class read a passage and had two "experts" rate the reading ability of each student. The scores provided by these two raters are then correlated, and the resulting correlation coefficient represents the interscorer reliability.

"The important issues are that training is often required and that a measure of the reliability of an evaluation of performance by raters is necessary."

Taqwa, Mandiri, Cendekia

uny.ac.id

REFERENCE

- Johnson, R. B., & Christensen, L. (2019). *Educational research: Quantitative, qualitative, and mixed approaches*. SAGE Publications, Incorporated.
- McMillan, J. H. (2017). *Classroom assessment: Principles and practice that enhance student learning and motivation*. USA: Pearson Education, Inc.
- Popham, W. J. (2018). *Classroom assessment: What teachers need to know*. University of California: Pearson Education, Inc.

Taqwa, Mandiri, Cendekia

uny.ac.id

THANK YOU

Taqwa, Mandiri, Cendekia

uny.ac.id

THE VALIDITY AND RELIABILITY OF THE TEST INSTRUMENT

Prof. Dr. Badrun Kartowagiran, M.Pd.

kartowagiran@uny.ac.id / badrun kartowagiran@gmail.com

Graduate School
Yogyakarta State University

Visiting Professor [13-22 October 2019]
Universiti Putra Malaysia, Malaysia

VALIDITY OF THE TEST INSTRUMENT

Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of the test

Validity is, therefore, the most fundamental consideration in developing and evaluating tests

Validity refers to the appropriateness or accuracy of the interpretations of the test scores

When the test scores are interpreted in multiple ways, each interpretation needs to be evaluated

Threats to Validity

"Validity is threatened when a test measures either less or more than the construct it is designed to measure"

Next

In addition to characteristic the test itself, factors external to the test can impact the validity of the interpretation of results.

Next

Linn and Gronlund (2000) identify numerous factors external to the test that can influence validity

- ☐ Instructional procedures
- ☐ Test administration and scoring procedures
- ☐ Students characteristic

Next

☐Instructional Procedures

With educational test, in addition to the content of the test influencing validity, the way the material is presented can influence validity.

For example, consider a test of critical thinking skills.
If the students were coached and given solution to the particular problems included on a test, validity would be compromised. This is a potential problem when teachers "teach the test"

Next

☐Test Administration and Scoring Procedures

Deviation from standard administrative and scoring procedures can undermine validity

In terms of administration, failure to provide the appropriate instructions or follow strict time limits can lower validity.
In terms of scoring, unreliable or biased scoring can lower validity

Next

☐Student Characteristics

Any personal factors that restrict or alter the examinees' responses in the testing situation can undermine validity

For example, if an examinee experiences high levels of test anxiety or is not motivated to put fort a reasonable effort, the results may be distorted

"Validity is not an all-or-none concept, but exist on a continuum"

Types of Validity

versus

Types of Validity Evidence

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

Types of Validity

- ❑ Content validity
- ❑ Criterion-related validity
- ❑ Construct validity

Types of Validity Evidence

- ❑ Validity evidence based on the test
- ❑ Validity evidence based on relations to other variables
- ❑ Validity evidence based on internal structure
- ❑ Validity evidence based on response processes
- ❑ Validity evidence based on consequences of testing

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

Types of Validity ❑ CONTEN VALIDITY

Content validity involves how adequately the test samples the content area of the identified construct

Is the content of the test relevant and representative the content domain?

Content validity is typically based on professional judgments about the appropriateness of the test content

Taqwa, Mandiri, Cendekia

uny.ac.id

Types of Validity ❑ CRITERION-RELATED VALIDITY

Criterion-related validity involves examining the relationships between the test and external variable that are thought to be direct measures of the construct

Studies of criterion-related validity empirically examine the relationships between test scores and criterion scores using correlation or regression analyses

Taqwa, Mandiri, Cendekia

uny.ac.id

Types of Validity ❑CONSTRUCT VALIDITY

Construct validity involves an integration of evidence that relates to the meaning or interpretation of test scores

This evidence can be collected using a wide variety of research strategies and design

Taqwa, Mandiri, Cendekia

uny.ac.id

Types of Validity Evidence ❑VALIDITY EVIDENCE BASED ON TEST CONTENT

Validity evidence based on the test content includes evidence derived from an analysis of the test content, which include the type of questions or tasks included in the test and administration and scoring guidelines

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

Types of Validity Evidence ❑VALIDITY EVIDENCE BASED ON RELATIONS TO OTHER VARIABLES

Validity evidence based on relations to other variables includes evidence based on an examination of the relationships between test performance and external variables of criteria

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

Types of Validity Evidence ❑VALIDITY EVIDENCE BASED ON INTERNAL STRUCTURE

Validity evidence based on internal structure includes evidence regarding relationships among test items and components

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

Types of Validity Evidence ❑VALIDITY EVIDENCE BASED ON RESPONSES PROCESSES

Validity evidence based on responses processes includes evidence derived from an analysis of the processes engaged in by the examinee or examiner

Next

Taqwa, Mandiri, Cendekia

uny.ac.id

Types of Validity Evidence ❑VALIDITY EVIDENCE BASED ON CONSEQUENCES OF TESTING

Validity evidence based on consequences of testing includes evidence based on an examination of the intended and unintended consequences of testing

Taqwa, Mandiri, Cendekia

uny.ac.id

"Sources of validity evidence differ in their importance according to factors such as the construct being measured, the intended use the test scores, and the population being assessed"

RELIABILITY OF THE TEST INSTRUMENT

Reliability can be defined as the proportion of test score variance due to true score differences

There are 5 major types of reliability according to Reynolds, C.R., Livingston, R. B., & Willson, V (2009)

Major types of Reliability

- Test-retest reliability
- Alternate forms reliability
- Split-half reliability
- Coefficient alpha and KR-20
- Inter-rater reliability

Test-Retest Reliability

Test-retest reliability is a method of estimating reliability that is exactly what its name implies.

The test is given twice and the correlation between the first set of scores and the second set of scores is determined

Next

Test-retest reliability is sensitive to measurement error due to time sampling and is an index of the stability of scores over time

One important consideration when calculating and evaluating test-retest reliability is the length of the interval between the two test administrations

The test-retest approach does have significant limitations, the most prominent being carryover effects from the first to second testing

Next

Example of Test-retest Reliability Computation

For example, suppose a math test given to six students on Monday is given again on the following Monday without any math having been taught in between these times. The six students make the following scores on the test:

Student	First Administration	Second Administration
	Score	Score
1	75	78
2	90	92
3	93	91
4	80	77
5	87	88
6	88	88

Next

"The correlation between the two sets of scores is 0.96. It could be concluded that this test is quite reliable"

Next

The main problem with test-retest reliability is that there is usually some memory or experience involved the second time the test is taken.

The scores may differ not only because of the unreliability of the test but also because the students themselves may have changed in some way.

For example:
Students may have gotten some answers correct on the retest by remembering or finding answers to some of the questions on the initial test

Next

"To determine test-retest reliability, the same test is administered twice to the same group of students, and their scores are correlated."

Generally, the longer the interval between test administrations, the lower the correlation.

Students can be expected to change with the passage of time, an especially long interval between testings will produce a "reliability" coefficient that is more a reflection of student changes on the attribute being measured than a reflection of the reliability of the test

Alternate-Form Reliability

- Based on simultaneous administration
- Based on delayed administration

Next

Taqwa, Mandiri, Cendekia uny.ac.id

Simultaneous Administration

Alternate-form reliability based on simultaneous administration is primarily sensitive to measurement error due to content sampling

Delayed Administration

Alternate-form reliability based on delayed administration is sensitive to measurement error due to content sampling and time sampling, but cannot differentiate the two types of errors

Taqwa, Mandiri, Cendekia uny.ac.id

Split-Half Reliability

Estimating **split-half reliability** involves administering a test and then dividing the test into two equivalent halves that are scored independently

The results on the first half of the test are then correlated with results on other half of the test by calculating the **Pearson Product Moment** correlation

Next

Taqwa, Mandiri, Cendekia uny.ac.id

There are many ways a test can be divided in half

For example,
one might correlate scores on the first half of the test with scores on the second half

➡ **But it is not good idea!**

WHY???

Next

Taqwa, Mandiri, Cendekia uny.ac.id

The previous example is not a good idea because the items on some test get more difficult as the test progresses, resulting in halves that are not actually equivalent

Other factors, such as practice effects, fatigue, or declining attention that increases as the test progresses, can also make the first and second halves of the test not equivalent

Next

Taqwa, Mandiri, Cendekia uny.ac.id

"A more acceptable approach would be to assign test items randomly to one half or other"

However, the most common approach is to use an odd-even split

Here all "odd"-numbered items go into one half and all "even"-numbered items go into the other half

A correlation is then calculated between scores on the odd-numbered and even-numbered items

Next

Taqwa, Mandiri, Cendekia uny.ac.id

The reliability coefficient does not take into account the reliability of the test when the two halves are combined

In essence, this initial coefficient reflects the reliability of only shortened, half-test

As a general rule, longer tests are more reliable than shorter test

Next

Taqwa, Mandiri, Cendekia uny.ac.id

Spearman-Brown Formula

To "**put the two halves of the test back together**" with regard to a reliability estimate, used a correction formula commonly referred to as the **Spearman-Brown Formula**

➡ **Reliability of Full Test:**

$$\frac{2\,x\,Reliability\,of\,Half\,Test}{1+Reliability\,of\,Half\,Test}$$

Next

Taqwa, Mandiri, Cendekia uny.ac.id

Spearman-Brown Formula

To "**put the two halves of the test back together**" with regard to a reliability estimate, used a correction formula commonly referred to as the **Spearman-Brown Formula**

➡ **Reliability of Full Test:**

$$\frac{2\,x\,Reliability\,of\,Half\,Test}{1+Reliability\,of\,Half\,Test}$$

Next

Taqwa, Mandiri, Cendekia uny.ac.id

Example of Spearman-Brown Formula Computation

Suppose the correlation between odd and even halves of your midterm in this course was **0.74**; the calculation using Spearman-Brown Formula would go as follows:

$$\text{Reliability} = \frac{2 \times 0.74}{1+0.74}$$

$$\text{Reliability} = \frac{1.48}{1.74}$$

$$\text{Reliability} = 0.85$$

➡ The reliability coefficient of 0.85 estimates the reliability of the full test when the odd-even halves correlated at 0.74. This demonstrates that the uncorrected split-half reliability coefficient presents an underestimate of the reliability of the full test

Next

Taqwa, Mandiri, Cendekia uny.ac.id

Coefficient Alpha Reliability and KR-20

Reliability estimates produced with these formulas can be thought of as the average of all possible split-half coefficient.

Next

KR-20 is applicable when test items that are scored dichotomously, that is, simply right or wrong, as 0 or 1

Coefficient alpha is a more general form of KR-20 that also deals with test items that produce scores with multiple values (e.g., 0, 1, or 2)

Next

KR-20 Formula

$$KR\text{-}20 = \frac{k}{k-1}\left(1 - \frac{\sum p_i x q_i}{SD^2}\right)$$

k = number of items
SD^2 = variance of total test scores
p_i = proportion of correct responses on item
q_i = proportion of incorrect responses on item

Next

Example of KR-20 Formula Computation

Item		p	q	pq
1		.40	.60	.24
2		.30	.70	.21
3		.60	.40	.24
4	$k = 6$	.70	.30	.21
5		.60	.40	.24
6		.30	.70	.21
				$\sum pq = 1.35$

$SD^2 = 4.00$

Next

$$KR\text{-}20 = \frac{k}{k-1}\left(1 - \frac{\sum p_i x q_i}{SD^2}\right)$$

$$KR\text{-}20 = \frac{6}{5}\left(1 - \frac{1.35}{4.00}\right)$$

$$KR\text{-}20 = 0.80$$

Coefficient Alpha Formula

$$\text{Coefficient alpha} = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum SD_i^2}{SD^2}\right)$$

k = number of items
SD_i^2 = variance of individual items
SD^2 = variance of total test scores

Next

Example of Coefficient Alpha Formula Computation

Suppose that examinees have been tested on four essay items in which possible scores range from 0 to 10 points. And $SD_1^2 = 9$; $SD_2^2 = 4.8$; $SD_3^2 = 10.2$; $SD_4^2 = 16$. If total score variance for the examination is $SD^2 = 100$, then what is the reliability?

$k = 4$ $SD_1^2 = 9$ $SD_2^2 = 4.8$ $SD_3^2 = 10.2$ $SD_4^2 = 16$ $SD^2 = 100$

$$\alpha = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum SD_i^2}{SD^2}\right)$$

$$\alpha = \left(\frac{4}{4-1}\right)\left(1 - \frac{9+4.8+10.2+16}{100}\right)$$

$$\alpha = 0.80$$

Inter-rater Reliability

Inter-rater reliability used if the scoring of a test relies on subjective judgment. It is important to evaluate the degree of agreement when different individuals score the test

Next

REFERENCE

"Linn, R. L., & Gronlund, N. E. (2000). Measurement and assessment in teaching (8 [th] ed.).

Reynolds, C. R., Livingston, R. B., & Willson, V. (2009). Measurement and assessment in education . Columbus, OH: Merrill."

THANK YOU

F. Foto-foto Kegiatan *Visiting Professor*

<table>
<tr>
<td>  </td>
<td>  </td>
</tr>
<tr>
<td>Gambar 1. Koordinasi Pelaksanaan Kegiatan Visiting Professor dengan Dekan di Universitas Mitra
(Prof. Badrun Kartowagiran menemui dan melakukan koordinasi dengan Dean of Faculty of Educational Studies)</td>
<td>Gambar 2. Discussion on Mutual Collaboration between UPM and UNY bersama dengan Kepala Bagian Internasional Universitas Mitra
(Prof. Badrun Kartowagiran menemui dan melakukan koordinasi bersama Kepala Bagian Internasional UPM terkait dengan mutual kolaborasi antara UNY dan UPM di masa yang akan datang)</td>
</tr>
<tr>
<td>  </td>
<td>  </td>
</tr>
<tr>
<td>Gambar 3. Kegiatan Lecturing pada Mata Kuliah Penilaian Kelas dengan Materi Authentic Assesment
(Prof. Dr. Badrun Kartowagiran memberikan perkuliahan pada mata kuliah Penilaian Kelas dengan menyampaikan materi tentang Penilaian Autentik)</td>
<td>Gambar 4. Kegiatan Lecturing pada Mahasiswa S1 dengan Materi tentang How to Develop effective Test
(Prof. Dr. Badrun Kartowagiran memberikan perkuliahan pada mahasiswa jenjang S1 di UPM. Materi yang disampaikan ialah tentang prosedur mengembangkan tes yang efektif)</td>
</tr>
</table>

Gambar 5. Kegiatan *Lecturing* pada Mahasiswa S3 dengan Materi tentang Validitas dan Reliabilitas
(Prof. Dr. Badrun Kartowagiran memberikan perkuliahan pada mahasiswa jenjang S3 di UPM. Materi yang disampaikan ialah Validitas dan Reliabilitas)

Gambar 6. Kegiatan Pelayanan konsultasi Mahasiswa
(Prof. Dr. Badrun Kartowagiran melayani mahasiswa S2 UPM yang melakukan konsultasi terkait cara penyusunan instruman

Gambar 7. Penerimaan Sertifikat Penghargaan dari Dekan *Faculty of Educational*