FORSCHEN.
FÖRDERN.
VERNETZEN.
DEUTSCHE GESELLSCHAFT FÜR FREMDSPRACHENFORSCHUNG

Universität Bremen

DGFF-Kolleg Sprachtests

23.03.2022

Claudia Harsch

harsch@uni-bremen.de

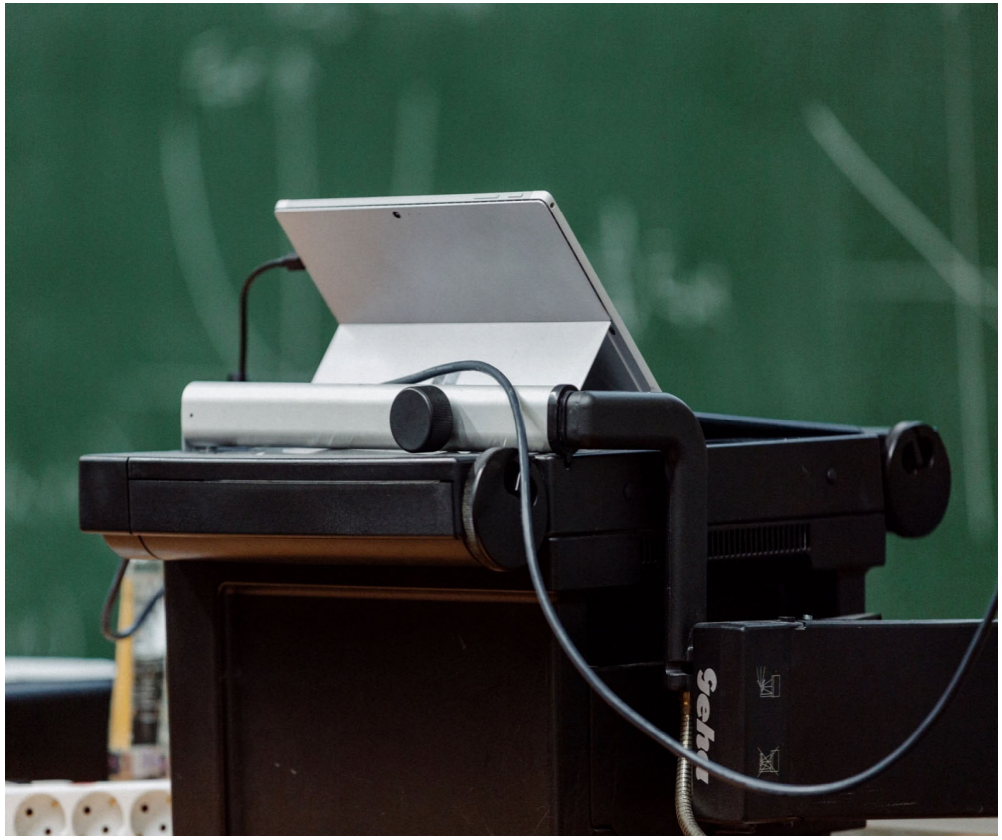


Foto: © Universität Bielefeld

DGFF Kolleg, 23. März 2022

Sprachtests in der Fremdsprachenforschung

Universität Bremen

Prof. Dr. Claudia Harsch
Universität Bremen

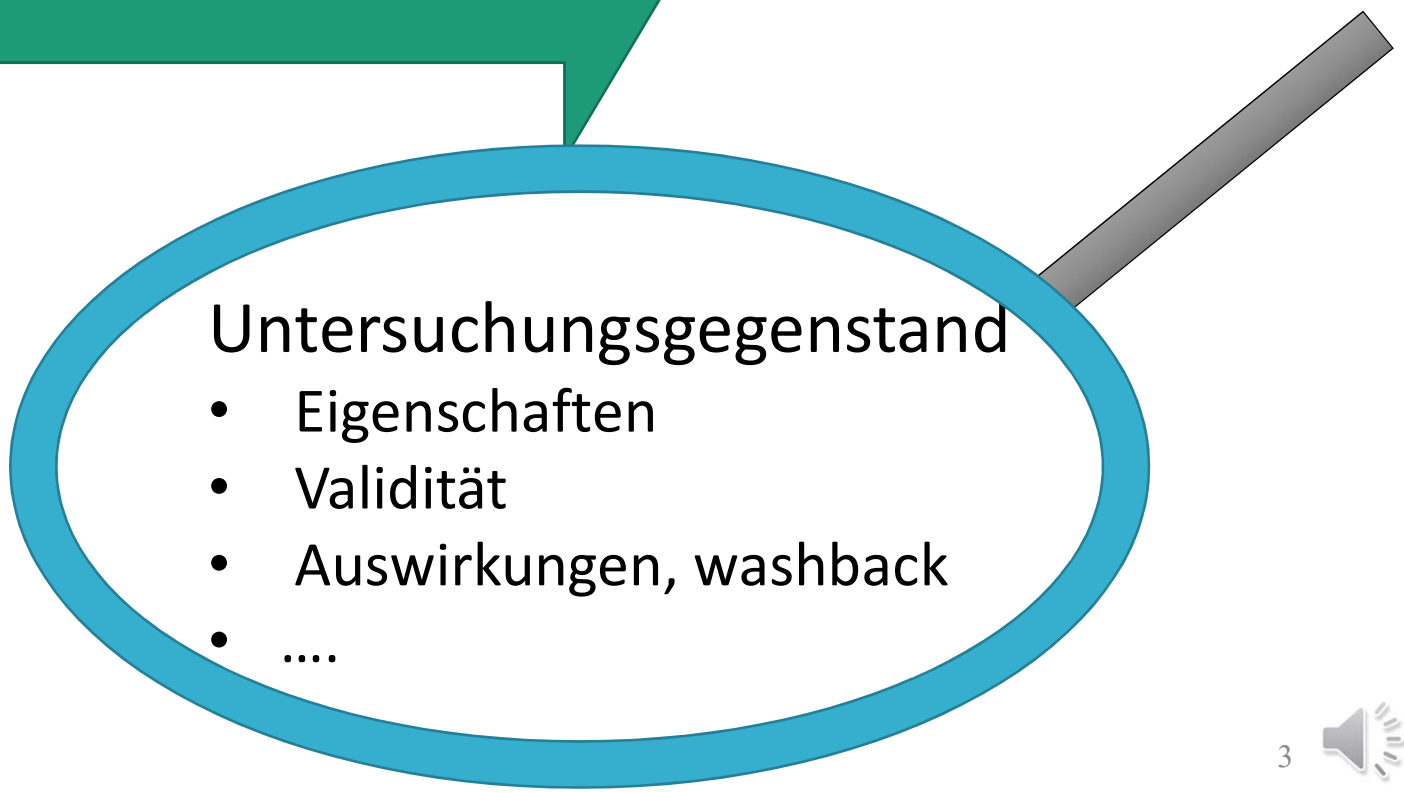
Was erwartet Sie?

1. Qualitätskriterien: Merkmale “guter” Testaufgaben
2. Grundlagen der Testanalyse: Validierung, Bildung von Kompetenzniveaus, Auswirkungen von Tests
3. Testeinsatz in Forschungsprojekten und Forschungsdesigns
4. Auswahl von geeigneten Tests
5. Diskussion: ethische Aspekte, Chancen und Grenzen

Forschungsinstrument oder Gegenstand?

Forschungsinstrument

- Designs
- Messung bestimmter Variablen
- ….



Eine Bitte vorab…

… bitte lesen Sie vorab die beiden Aufsätze, die ich Ihnen zur Verfügung gestellt habe:

Harsch, C. (2012). Der Einsatz von Sprachtests in der Fremdsprachenforschung: Tests als Untersuchungsgegenstand und Forschungsinstrument. In: Doff, Sabine (Hg.). *Fremdsprachenunterricht empirisch erforschen: Grundlagen, Methoden, Anwendung*. Tübingen: Narr, 150 – 183.

Harsch, C. (2016, Neuauflage im Druck). Testen. In: Daniela Caspari, Friederike Klippel, Michael Legutke & Karen Schramm (Hrsg.). *Forschungsmethoden in der Fremdsprachendidaktik. Ein Handbuch*. Tübingen: Narr, 204-217.

Bitte die pdfs vertraulich behandeln und nicht weitergeben!

Begrifflichkeiten

BITTE SCHLAGEN SIE DIE FOLGENDEN BEGRIFFLICHKEITEN NACH

- Test – Prüfung – Bewertung – Beurteilung – Evaluation – Assessment
- Aufgabe – Task – Input – Prompt – Stimulus – Item – Frage – Antwort – Lösung
- summativ – formativ
- Kompetenz/Lernstand (*proficiency*) – Lernerfolg/Fortschritt (*achievement*) – Diagnose – Einstufung (*placement*)
- Kompetenz – Performanz – Leistung

1. Merkmale „guter“ Testaufgaben

Gütekriterien in der Aufgabenentwicklung

Objektivität

Unabhängigkeit der Beurteilung von den Beurteilenden oder dem Beurteilungs-instrument

Reliabilität

Zuverlässigkeit der Aufgabe oder der Bewertung, Messkonsistenz, i.d.R. statistische Verfahren zur Überprüfung

Validität

Misst die Aufgabe, was sie messen soll? Auch: Hat sie die intendierten Konsequenzen? Statistische, aber auch qualitative Verfahren zum Sammeln von Validitätsevidenzen

Praktikabilität

Ressourcen, Umsetzbarkeit, Machbarkeit: Testentwicklung, Administration, Aufgaben

Ethische Aspekte

Inhalte, Durchführung, Konsequenzen

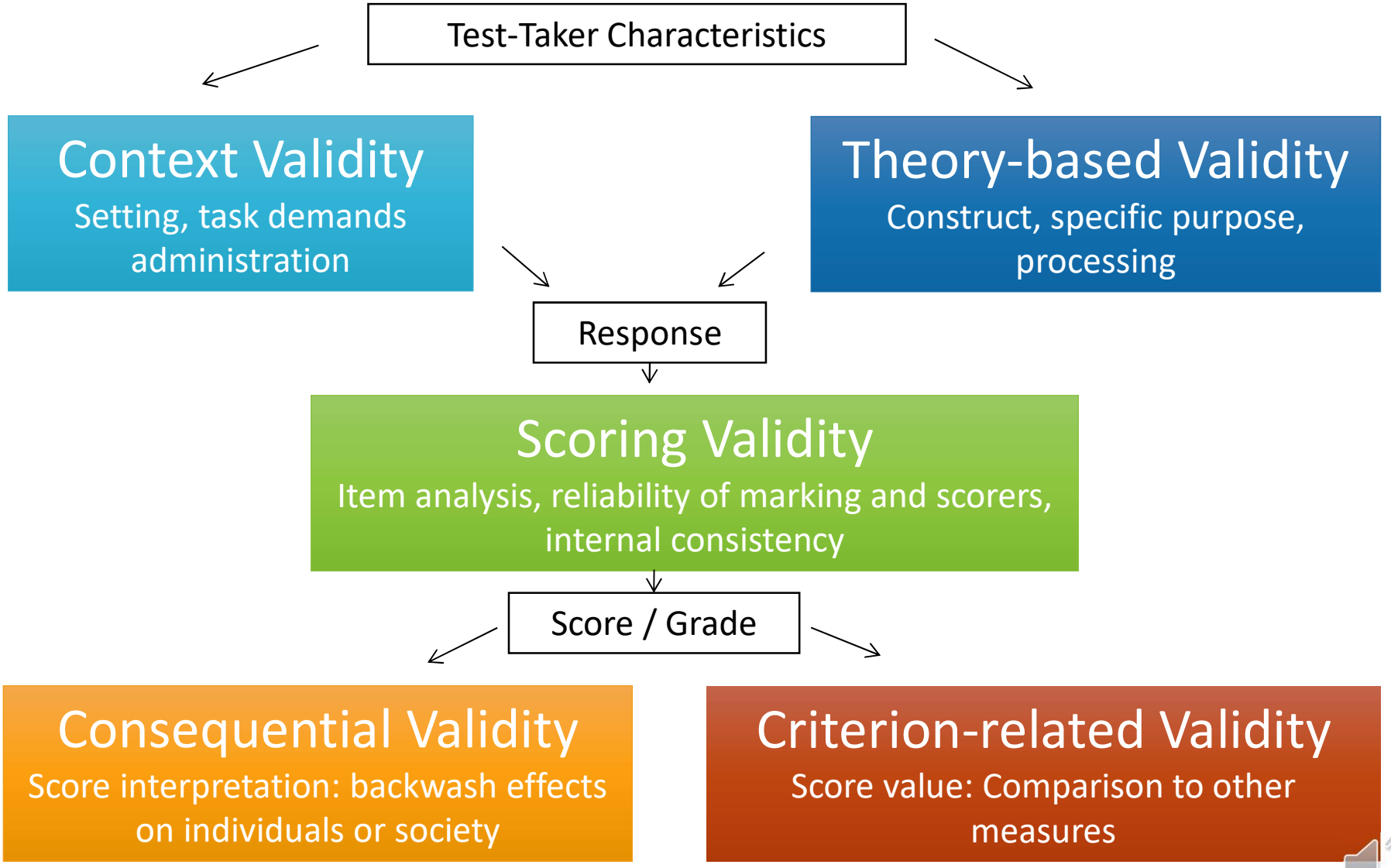
Kritik am Sprachtesten

- Objektivität: standardisiertes Testen lasse keine subjektiven Interpretationen und Lösungswege zu
- Reliabilität: Lernzuwachs, Erinnerungseffekte seien unerwünscht
- Validität: Individuelle Momentaufnahmen ≠ Kommunikation im Leben (interaktiv, ko-konstruktiv)

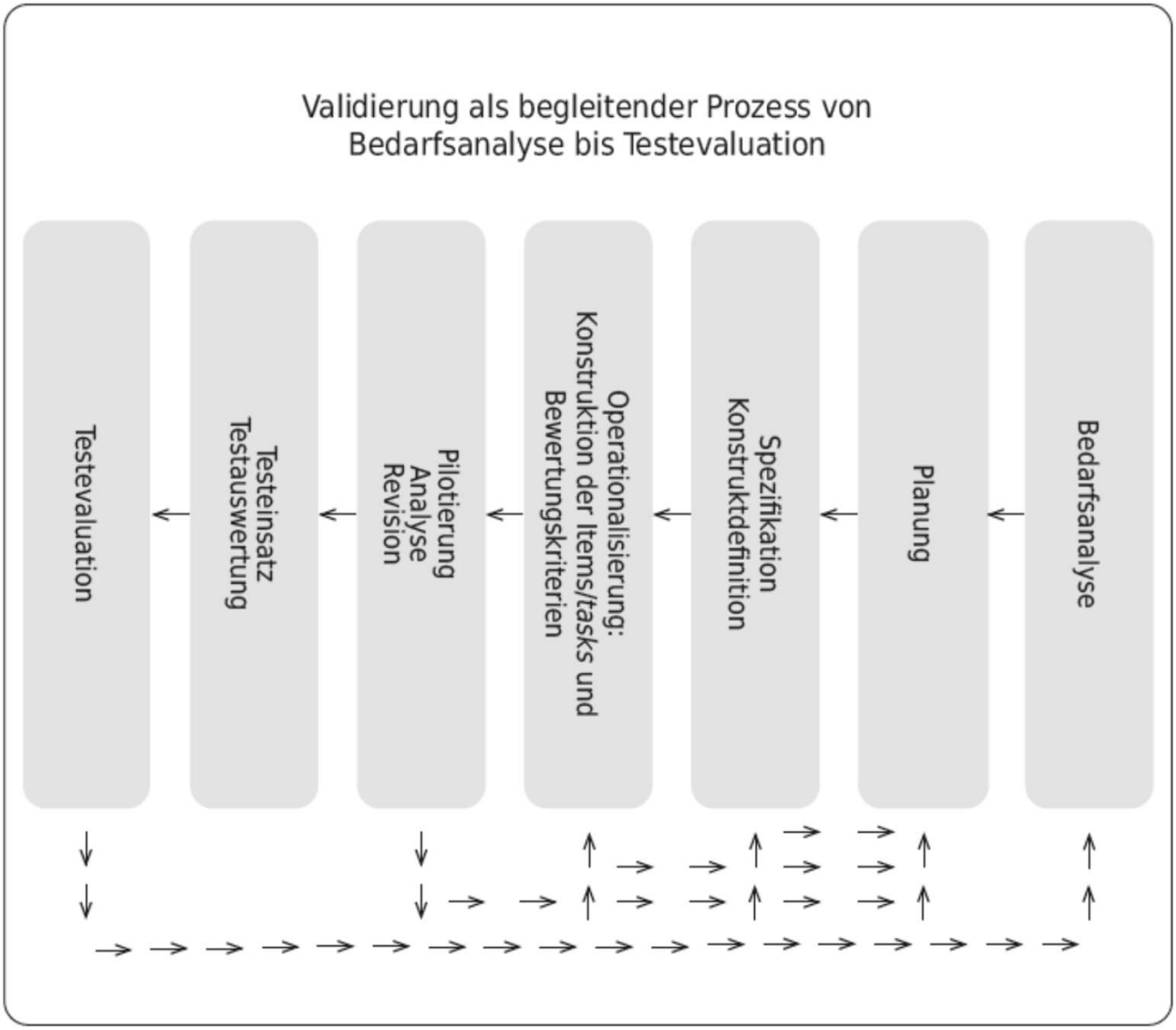
- Alternative Herangehensweisen
  dynamic assessment (Poehner 2008), interaktives Beurteilen (Ahmed & Pollitt 2010), *learning-oriented assessment* (z.B. Assessment Reform Group 2002)

2. Grundlagen der Testanalyse: Validierung, Bildung von Kompetenzniveaus, Auswirkungen von Tests

Weir's Cognitive Validity Framework (2005: 44ff)



Testaufgabenentwicklung

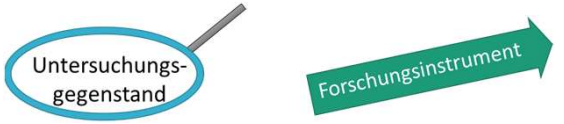


[11] Test development Manual der Association of Language Testers in Europe (2011).

Vorgehen: Zyklischer Prozess

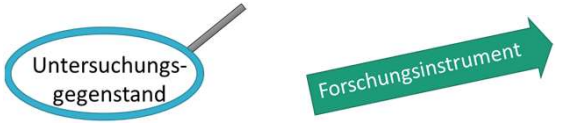
- Konstrukt: Theoretische Vorstellung, was die Aufgabe messen soll
  - Basis: z.B. Theorien, Bildungsstandards, Curricula, GER
  - Definition in den Test Specifications
- Operationalisierung: valide Umsetzung des Konstrukts in Au[illegible]n
  - Qualitative Verfahren der Testentwicklung (zB text mapping, AG)
  - Qualitative Verfahren der Testcharakterisierung (Einschätzung durch Entwickler:innen / Expert:innen, z. B. Dutch Grid, Grids des Europarats
- Erprobung(en), Analyse, Überarbeitung(en)
  - Quantitative Analysen der Aufgaben / empirischen Datenbasis
  - Qualitative Analysen der Antworten / Performanzen der Lernenden
  - Prozessanalysen (TAP, eye-tracking)
- Einsatz und Auswertung
- Anbindung an Kompetenzniveaus / GER
  - Standard-setting Verfahren
- Auswirkungen, washback

Qualitative Analysen



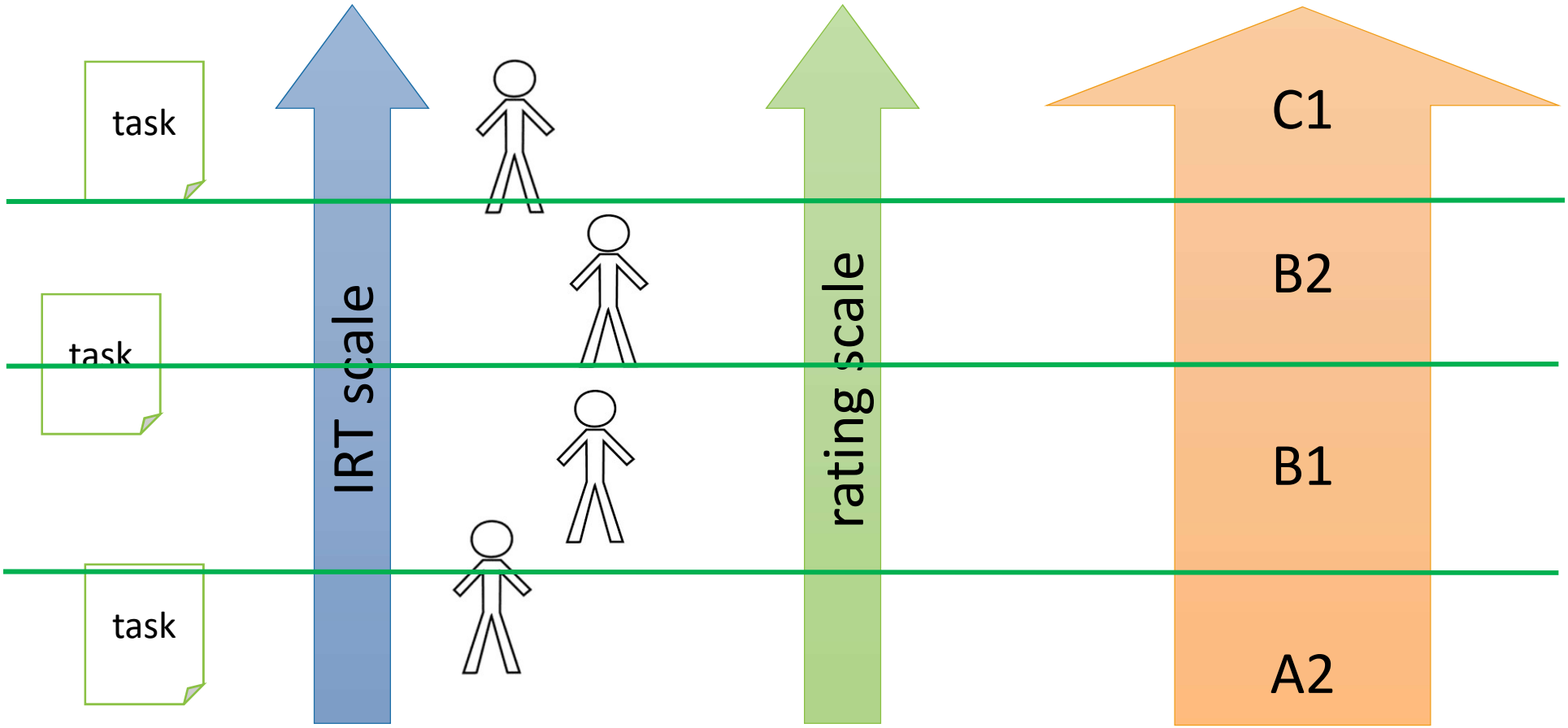
- Aufgaben-/Inhaltsanalysen: Expert:innenurteile
- Diskursanalysen
  - Test discourse ≠ echte Kommunikation
  - Effekte von tasks, test takers, interlocutors
  - Test comparison
- Prozessanalysen
  - Fokus auf test takers, interlocutors, raters
  - Introspektion: concurrent (think-aloud) or retrospective approaches
  - Eye tracking
- Kontextanalysen zur Auswirkung
  - Ethnographische Methoden, e.g. Interview, Beobachtung, Fragebogen
  - Prädiktive Validität (oft mixed-methods)

Quantitative Analysen



- Klassische Testtheorie, Testgüte-Indikatoren
  - Lösungshäufigkeit
  - Reliabilität, interne Konsistenz
  - Diskriminanzanalysen
  - Distraktorenanalysen
- Ähnlichkeiten: Korrelationen
  - Testvergleich
  - Rating Reliabilität
  - Dimensionsbestimmung (Faktorenanalysen)
- Unterschiede
  - ANOVA, t-tests, Chi-square
  - Item bias: DIF
- IRT, Probabilistische Testtheorie
  - Rasch, FACETS

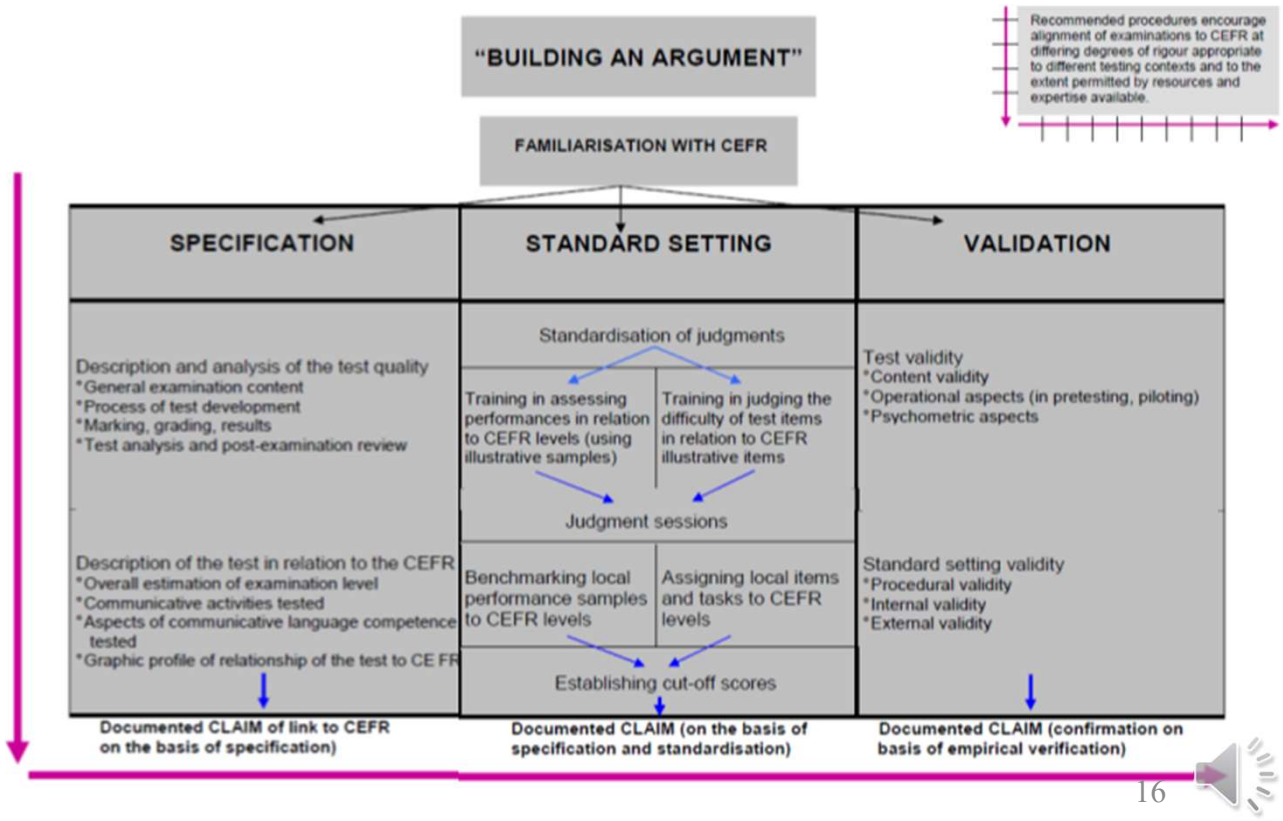
Standard Setting – Bildung von Kompetenzniveaus



Standard Setting – Bildung von Kompetenzniveaus

Manual for relating Language Examinations to the CEFR:

- erläutert mögliche Schritte und Prozeduren;
- regt zu transparenter Dokumentation durch Testerstellende
- stellt praktisch Verfügung.



Standard Setting – Bildung von Kompetenzniveaus

CASE STUDY 3: GERMANY

Standard Setting in the context of evaluating the German Educational Standards in EFL

Claudia Harsch

www.ecml.at/Events/ECMLcolloquium7December2016/Webstream/tabid/2986/language/en-GB/Default.aspx

video at https://youtu.be/GQL2IqRtkW

Testentwicklung- und Analyse: Aus- und Fortbildung

Es gibt *pre-conference workshops* und Kurse, die gezielt in quantitativen und qualitativen Methoden der Testentwicklung, -auswertung und -analyse schulen, e.g.

- EALTA pre-conference workshops
- LTRC pre-conference workshops
- Workshops der ALTE
- Sommerkurse, z.B. in Lancaster

3. Tests in Forschungsprojekten; Forschungsdesigns

Einsatz in der Fremdsprachenforschung

a) Interventionsstudien

- Effekte von Interventionen
- (quasi-)experimentelle Designs – **s. Zusatzaufgabe**
- Leistungsunterschiede vor/nach Intervention bzw. zwischen Gruppen
- Bei Veränderungsmessung: prä-/post-Test Designs

➢ vergleichbare, kalibrierte Tests, aber nicht dieselben

➢ ggf. Einstufungstests zur Gruppeneinteilung

- Flankierende qualitative Instrumente

b) Standardisierte Leistungsmessung

- *Large-scale assessment* Studien wie PISA, DESI, Ländervergleich
- Generalisierbarkeit, Bildungsmonitoring

Experimentelle Designs

- Experimentelle Designs:
  - eine unabhängige Variable wird manipuliert, um Effekt auf abhängige Variable zu untersuchen
  - Kontrollgruppe ohne Manipulation, experimentelle Gruppe(n) mit Manipulation(en)
  - zufällige Zuordnung der Teilnehmenden in die beiden Gruppen
- In unserem Feld: Nicht immer lassen sich alle Bedingungen eines echten Experiments umsetzen (z.B. Schulklassen, Ethik) – deshalb mehrere Vorstufen:
  - Quasi-experimentell: Kontrollgruppe, aber keine zufällige Zuordnung, sondern Beobachtung in "natürlichen" Gruppen (zB Klassenverband)
  - Prä-experimentell: keine Kontrollgruppe
- Ex-post-facto Design, z.B. Korrelationsstudien: Beziehungen zwischen Variablen, ohne nach Ursachen zu suchen; keine Manipulation, keine Kontrollgruppe notwendig

Experimentelle Designs

- Experimentelle Designs zum Testen von Hypothesen:
  - eine unabhängige Variable wird manipuliert, um Effekt auf abhängige Variable zu untersuchen
  - Kontrollgruppe ohne Manipulatio
    Manipulation
  - zufällige Zuordnung der Teilnehm
- In unserem Feld: Nicht immer la
  eines echten Experiments umse
  deshalb mehrere Vorstufen:
  - Quasi-experimentell: Kontrollgrup ng,
    sondern Beobachtung in "natürlic nd)
  - Prä-experimentel: keine Kontrollg
- Ex-post-facto Design, z.B. Ko n
  zwischen Variablen, ohne nach Ursachen zu suchen, keine
  Manipulation, keine Kontrollgruppe notwendig

Zusatzaufgabe:
s. Handout zu den drei Interventionsstudien und ihren jeweiligen Designs

4. Auswahl existenter Instrumente

Vorabprüfung

- Gibt es für mein Forschungsziel, meine Fragen, meine Zielgruppe, mein Design etc. bereits ein Instrument, das ich einsetzen oder adaptieren könnte?
- Evaluation der existenten Instrumente
  s. Handout und Zusatzaufgabe

- Ist ein passendes Instrument gefunden:
  Kontaktaufnahme und Klärung der Nutzungsrechte
- Gibt es kein passendes Instrument:
  s. 2 oben, Testentwicklung/Validierung…

Praxis: Testeinsatz in der eigenen Studie

- Einverständniseinholung, bei Schulen ggf. über Ministerium, Schulleitungen, Eltern
- Vorbereitung der Testmaterialien (print oder online Umgebung), Audio-Abspielgeräte klären
- ggf. Schulung Testleiter:innen (Skript)
- ggf. Aufnahme mündlicher Prüfungen, Schulung interlocutors (guide)
- ggf. Bewerter:innen-Training (Rating Scale, Benchmark-Leistungen)
- Planung Auswertung, Dateneingabe, cleaning, Analysen
- Zeitnahe Rückmeldung an Schüler:innen und Lehrer:innen, Aufbereitung der Ergebnisse entsprechend der Zielgruppe

5. Diskussion im Seminar: Ethische Aspekte, Chancen und Grenzen

Vorüberlegungen fürs Seminar

- Wo sehen Sie die Grenzen des Testeinsatzes in der Fremdsprachenforschung?
- Was können Tests in der Forschung leisten, was nicht?
- Wo sehen Sie ethische Aspekte, die es zu bedenken gibt?
  Vgl. auch ALTE (2020), EALTA (2006), ILTA (2018) guidelines

Bibliographie

Ahmed, A., & Pollitt, A. (2010). The Support Model for interactive assessment. *Assessment in Education: Principles, Policy & Practice, 17*(2), 133 – 167.

Assessment Reform Group (2002). *Assessment for Learning: 10 Principles*. Cambridge: University School of Education.

Harsch, C. (2019). What it means to be at a CEFR level. Or why my Mojito is not your Mojito – on the significance of sharing Mojito recipes. In Ari Huhta, Neus Figueras & Gudrun Erickson (eds). *Developments in language education – a memorial volume in honour of Sauli Takala*, 76-93. EALTA / University of Jyväskylä, Finland. Available online: http://www.ealta.eu.org/documents/resources/Developments%20in%20Language%20Education%20A%20Memorial%20Volume%20in%20Honour%20of%20Sauli%20Takala.pdf.

Harsch, C. (2012). Der Einsatz von Sprachtests in der Fremdsprachenforschung: Tests als Untersuchungsgegenstand und Forschungsinstrument. In: Doff, Sabine (Hg.). *Fremdsprachenunterricht empirisch erforschen: Grundlagen, Methoden, Anwendung.* Tübingen: Narr, 150 – 183.

Harsch, C. (2016, Neuauflage im Druck). Testen. In: Daniela Caspari, Friederike Klippel, Michael Legutke & Karen Schramm (Hrsg.). *Forschungsmethoden in der Fremdsprachendidaktik. Ein Handbuch*. Tübingen: Narr, 204-217.

Poehner, M. (2008*). Dynamic assessment: a Vygotskian approach to understanding and promoting L2 development.* Berlin: Springer.

Weir, C. J. (2005). *Language Testing and Validation*. Oxford: Palgrave.

URLs

Association of Language Testers in Europe (ALTE) *Manual for Language Test Development and Examining* (2011): https://www.alte.org/resources/Documents/ManualLanguageTest-Alte2011_EN.pdf

ALTE *Principles of good practice* (2020): https://www.alte.org/resources/Documents/ALTE%20Principles%20of%20Good%20Practice%20Online%20(Final).pdf

Council of Europe *Grids to specify language tests* (n.d.): https://www.coe.int/en/web/common-european-framework-reference-languages/relating-examinations-to-the-cefr

Dutch Grid: https://www.lancaster.ac.uk/fss/projects/grid/

EALTA *Guidelines* (2006): http://www.ealta.eu.org/guidelines.htm

ILTA *Code of Ethics* (2018): https://www.iltaonline.com/page/CodeofEthics

Standard Setting example for German Educational Standards: www.ecml.at/Events/ECMLcolloquium7December2016/Webstream/tabid/2986/language/en-GB/Default.aspx
video at https://youtu.be/GQL2IqRtkWE

Weiterführende Bibliographie

Bachman, Lyle F. (2004). *Statistical analyses for language assessment.* Cambridge: Cambridge University Press.

Bachman, Lyle F./Kunnan, Anthony J. (2005). *Statistical analyses for language assessment. Workbook and CD.* Cambridge: Cambridge University Press.

Diese beiden Bände geben eine sehr gute Einführung in die statistische Testanalyse. Das Workbook (mit CD) ergänzt die Monographie um praktische Beispiele und Übungen an realen Datensätzen.

Bachman, Lyle F./Palmer, Adrian S. (2010*). Language assessment in practice.* Oxford: Oxford University Press.

Diese Monographie ist eines der Standardwerke in der Testliteratur; sie diskutiert alle wesentlichen Aspekte des Designs, der Entwicklung und des Nutzens von Sprachtestes und Sprachbeurteilung. Insbesondere die Ausführungen zum *Assessment Use Argument* sind bemerkenswert, da sie die Testnutzung und den Einsatz von Beurteilung in den Mittelpunkt rücken.

Douglas, Dan (2010). *Understanding Language Testing.* London: Hodder Education.

Dieser Band bietet eine kurze und leicht verständliche Einführung in die Natur, Entwicklung, Analyse und den Einsatz von Sprachtests.

Hinger, Barbara/Stadler, Wolfgang (2018). *Testen und Bewerten fremdsprachlicher Kompetenzen*. Tübingen: Narr.

Dieses Studienbuch wendet sich sowohl an Studierende als auch PraktikerInnen und gibt einen guten Einblick in die relevanten Aspekte des Testens und Bewertens. Dabei werden der aktuelle Forschungsstand und praktische Anwendungsbeispiele mit einbezogen.

Vielen Dank für Ihre Aufmerksamkeit
Ich freue mich auf unser Seminar

Gerne dürfen Sie mir Ihre Fragestellungen, Forschungs-Designs etc. vorab zukommen lassen:

harsch@uni-bremen.de

Universität Bremen