

PAL Test Entwicklungsupdate zur Itemrevision

Zwischenstand der systematischen Itemrevision

Stand: 18. September 2026

Kurzfassung

Die Pilotierung des PAL-Tests V4.0 ist in die Phase der systematischen Itemrevision übergegangen.

In einem ersten Revisionsblock wurden die 15 bislang am häufigsten exponierten Items untersucht. Neben Schwierigkeitsindex und korrigierter Trennschärfe wurden insbesondere die Auswahlhäufigkeiten der Antwortalternativen sowie sprachliche Merkmale der Distraktoren geprüft. Auch ihre fachsystematische und inhaltliche Gestaltung wurde analysiert.

Ein zentraler Zwischenbefund betrifft die Funktion der falschen Antwortalternativen. Bei mehreren Items wurden einzelne Distraktoren selten oder überhaupt nicht gewählt. Als mögliche Ursachen zeigten sich sprachliche Lösungshinweise und unterschiedliche Abstraktionsebenen. Auch fachsystematisch auffällige Antwortgruppen und wenig plausible Alternativen kamen als Erklärung infrage.

Bei einem anderen Item wurde die richtige Alternative in den vollständig vorliegenden Antwortdetails von 19/19 Bearbeitern erkannt. Dennoch lag der Schwierigkeitsindex bei $p = .10$. Ein besonders attraktiver Distraktor wurde von 15/19 zusätzlich gewählt. Die Konstruktion der Antwortalternativen dürfte die beobachtete Schwierigkeit daher wesentlich beeinflusst haben.

Aus der Analyse wurden folgende Revisionsprinzipien abgeleitet.

- sprachliche und syntaktische Parallelität der Antwortalternativen,
- fachsystematische Vergleichbarkeit der Distraktoren,
- wissenschaftlich und terminologisch belastbare Antwortalternativen,
- Vermeidung unbeabsichtigter Lösungshinweise und Mehrdeutigkeiten,
- erneute Prüfung der AFB-Zuordnungen,
- Rohdatenprüfung von SI und korrigierter Trennschärfe vor der Interpretation.

Die Revision soll die Eindeutigkeit und Inhaltsvalidität der Items verbessern. Sie soll außerdem deren diagnostische Funktion stärken. Eine künstliche Erhöhung der Schwierigkeit würde diesem Ziel widersprechen.

Substanziell revidierte Items werden anschließend in einer zweiten Pilotierungsphase erneut empirisch geprüft.

Pilotierung → Analyse → Revision → Repilotierung

Langfassung

PAL-Test V4.0 – Zwischenstand der systematischen Itemrevision

Die Pilotierung des PAL-Tests V4.0 ist inzwischen in eine erste systematische Revisionsphase übergegangen. Grundlage hierfür bilden die bislang erhobenen Testdurchläufe sowie die daraus gewonnenen Item- und Antwortdaten.

Für einen ersten vertieften Analyseschritt wurden die 15 am häufigsten exponierten Items ausgewählt. Die Exposition dieser Items liegt derzeit zwischen $n = 20$ und $n = 42$.

Die Analyse verbindet den Schwierigkeitsindex (p) und die korrigierte Item-Restscore-Korrelation mit einer qualitativen Prüfung der Antwortalternativen. Berücksichtigt werden insbesondere die Auswahlhäufigkeiten der Distraktoren, sprachliche und syntaktische Merkmale, fachsystematische Vergleichbarkeit, mögliche Lösungshinweise, die Passung zum intendierten Anforderungsbereich sowie die berufliche Plausibilität der Antwortoptionen. Bearbeitungszeiten werden ergänzend und aufgrund ihrer Verteilung zunächst ausschließlich explorativ betrachtet.

Distraktorfunktion als zentraler Zwischenbefund

Bei mehreren der bislang analysierten Items zeigen einzelne falsche Antwortalternativen eine geringe oder keine erkennbare Distraktorfunktion. Teilweise wurden Distraktoren trotz einer ausreichenden Zahl von Itemexpositionen überhaupt nicht ausgewählt.

Die qualitative Analyse weist dabei auf wiederkehrende Konstruktionsmuster hin. Dazu gehören insbesondere sprachlich auffällige Formulierungen, absolute oder stark einschränkende Begriffe, Unterschiede in Satzstruktur und Komplexität, unterschiedliche fachsystematische Ebenen sowie Antwortalternativen, die gegenüber der richtigen Lösung inhaltlich zu offensichtlich falsch erscheinen.

Distraktoren sollten fachlich plausibel sein und etablierte Begriffe verwenden. Eine Alternative kann in einem anderen Zusammenhang angemessen sein. Bezogen auf die konkrete Fragestellung muss sie jedoch eindeutig falsch sein.

Die Attraktivität eines Distraktors soll aus seiner fachlichen Plausibilität entstehen. Sprachliche Täuschung und künstliche Komplexität würden die beabsichtigte Beurteilung verzerren.

Empirische Schwierigkeit und inhaltliche Anforderung unterscheiden

Ein besonders aufschlussreicher Befund ergab sich bei einem Multiple-Choice-Item mit einem Schwierigkeitsindex von lediglich $p = .10$. Eine isolierte Betrachtung des Kennwertes würde zunächst auf ein sehr schwieriges Item hindeuten.

Die Analyse der Antwortalternativen zeigt jedoch ein differenzierteres Bild: In den vollständig vorliegenden Antwortdetails wurde die richtige Alternative von 19/19 Bearbeitern ausgewählt. Gleichzeitig wurde ein bestimmter Distraktor von 15/19 Bearbeitern zusätzlich markiert.

Die geringe Quote vollständig richtiger Lösungen dürfte damit wesentlich durch die hohe Attraktivität beziehungsweise mögliche Mehrdeutigkeit dieses Distraktors beeinflusst worden sein. Die richtige inhaltliche Lösung wurde dagegen nahezu durchgängig erkannt.

Der Befund verdeutlicht die begrenzte Aussagekraft eines isoliert betrachteten Schwierigkeitsindex bei Multiple-Choice-Items. Die Interpretation eines niedrigen p -Wertes erfordert eine Analyse des konkreten Antwortverhaltens.

Vorsicht bei der Interpretation der Trennschärfe

Auch die korrigierte Item-Restscore-Korrelation wird deshalb für jedes analysierte Item unmittelbar anhand der bereinigten Rohdaten überprüft.

Bei Items mit sehr geringer Varianz ist besondere Vorsicht erforderlich. Wird eine Aufgabe bei 20 Bearbeitungen lediglich ein- oder zweimal falsch beantwortet, können einzelne Testdurchläufe die resultierende Trennschärfe stark beeinflussen. Auch ein numerisch hoher Wert bedarf unter diesen Bedingungen einer Prüfung seiner Stabilität.

Quantitative Kennwerte und qualitative Itemanalyse werden daher bewusst miteinander verbunden.

Konsequenzen für die Itemrevision

Aus dem ersten Revisionsblock wurden mehrere wiederkehrende Kriterien für die weitere Überarbeitung des Itempools abgeleitet:

- sprachliche und syntaktische Parallelisierung der Antwortalternativen,
- Vermeidung auffälliger Signalwörter und unbeabsichtigter Lösungshinweise,
- Verwendung wissenschaftlich etablierter und terminologisch korrekter Distraktoren,
- stärkere fachsystematische Homogenität innerhalb eines Items,
- Vermeidung konstruktirrelevanter Informationen im Fragenstamm,
- Überprüfung schwacher oder potenziell mehrdeutiger Distraktoren,
- Prüfung auffälliger Antwortstrukturen, insbesondere bei MC-Items mit vier richtigen und nur einer falschen Alternative,
- gegebenenfalls Umstellung einzelner SC- auf MC-Formate,
- erneute Überprüfung der Zuordnung zu den Anforderungsbereichen.

Anforderungsbereich und empirische Itemschwierigkeit werden getrennt betrachtet. Der Anforderungsbereich richtet sich nach der erforderlichen kognitiven Leistung. Die sprachliche Komplexität ist dafür allein kein Maßstab. Auch ein empirisch leichtes Item kann eine fachlich relevante Grundlage valide und eindeutig erfassen.

Erneute Pilotierung

Der erste Revisionsblock der 15 am häufigsten exponierten Items ist weitgehend abgeschlossen. Bei einzelnen Items steht die abschließende Konstruktion geeigneter Distraktoren noch aus.

Substanziell veränderte Items gelten psychometrisch als neue Fassungen. Ihre Kennwerte müssen deshalb in einer erneuten Pilotierung bestimmt werden.

Nach Abschluss der Revision folgt eine zweite Pilotierungsphase. Die erneute Datenerhebung soll zeigen, wie sich die Distraktorfunktion verändert hat. Die Schwierigkeitswerte werden erneut beurteilt. Auch die Stabilität der Trennschärfen ist zu prüfen.

Pilotierung → Rohdatenanalyse → quantitative und qualitative Itemanalyse → Revision → Repilotierung

Die Pilotierung untersucht einzelne Items und liefert zugleich empirische Hinweise auf die Qualität der Itemkonstruktion. Auch Eigenschaften des Messverfahrens werden dadurch einer Überprüfung zugänglich.