

PAL Test Zwischenstandsprotokoll V3

Pilotierung, Itemexposition, Itemanalyse und Bearbeitungszeiten

Stand: 11. September 2026

Kurzfassung

Ausgangslage

Der PAL-Test V4.0 umfasst einen aktiven Pool von 120 Single-Choice- und Multiple-Choice-Items. Zum aktuellen Auswertungszeitpunkt lagen 99 registrierte Testdurchläufe vor. Durchläufe, deren Pseudonym die Zeichenfolge „test“ enthielt und die damit eigenen beziehungsweise technischen Erprobungen zugeordnet werden konnten, wurden ausgeschlossen. Es verblieben 78 reguläre Durchläufe mit 1.350 auswertbaren Itembearbeitungen. Alle 120 Items wurden mindestens einmal präsentiert.

Itemexposition und Ziehungslogik

Die Exposition der Items war deutlich ungleich verteilt und reichte von $n = 2$ bis $n = 42$ bei einem Median von 9,5 Bearbeitungen. Beispielsweise wurde PAL095 bislang 42-mal, PAL019 dagegen nur zweimal präsentiert. Die technische Prüfung zeigte, dass die Randomisierung grundsätzlich funktionierte, die bisherige Auswahlroutine jedoch Themenkategorien und zusätzlich das Verhältnis von Single-Choice- und Multiple-Choice-Items balancierte. Da die Kategorien unterschiedlich große Itempools enthalten, entstanden systematisch unterschiedliche Auswahlwahrscheinlichkeiten einzelner Items. Für die psychometrische Pilotierung ist dies ungünstig, weil itembezogene Kennwerte auf sehr unterschiedlichen Beobachtungszahlen beruhen.

Konsequenz für die Auswahlroutine

Die 120 Items bleiben zunächst inhaltlich unverändert. Angepasst wird ausschließlich die Ziehungslogik. Für die weitere Pilotierung soll die Itemexposition stärker berücksichtigt werden. Selten präsentierte Items erhalten eine höhere Auswahlwahrscheinlichkeit, stark exponierte Items werden vorübergehend geringer gewichtet. Die inhaltliche Mischung sowie eine angemessene Berücksichtigung von SC- und MC-Items sollen erhalten bleiben, dürfen aber nicht erneut zu ausgeprägten systematischen Expositionsunterschieden führen. Zeitpunkt und Art der technischen Revision werden dokumentiert.

Erste Itemanalysen

Erste Itemanalysen sind möglich, sofern die jeweilige Beobachtungszahl transparent berücksichtigt wird. Als interne Arbeitsregel werden Items mit $n < 10$ zunächst beobachtet. Bei $n = 10-19$ erfolgt ein orientierendes Screening. Ab etwa $n \geq 20$ ist eine erste vertiefte Zwischenanalyse vorgesehen. Diese Schwellen dienen der vorläufigen Arbeitsorganisation. Revisionsentscheidungen erfordern eine gemeinsame Beurteilung der Itemschwierigkeit und korrigierten Trennschärfe. Antwortmuster und Bearbeitungszeiten werden ergänzend berücksichtigt. Soweit sinnvoll, wird das subjektive Schwierigkeitsrating einbezogen.

Bearbeitungszeiten und Extremwerte

Die Analyse von PAL095 verdeutlicht den Bedarf an einer Prüfung der Zeitmessung. Für das Item lagen 42 reguläre Bearbeitungen vor. Ein Zeitwert betrug rund 612,7 Sekunden. Das Item stand in diesem Durchlauf an Position 1 von 20. Die gespeicherten Interaktionsdaten lagen deutlich früher innerhalb des registrierten Zeitfensters. Auch der zweitlängste Wert von rund 105,0 Sekunden trat bei Position 7 von 10 auf. Die Daten begründen Zweifel daran, dass die gespeicherte Itemzeit in jedem Fall aktive kognitive Bearbeitung abbildet.

Statistischer Umgang mit Zeitdaten

Bearbeitungszeiten werden deshalb zunächst unbereinigt mit n , arithmetischem Mittel, Standardabweichung, Median, Quartilen beziehungsweise Interquartilsabstand sowie Minimum und Maximum beschrieben. Median und IQR ergänzen Mittelwert und Standardabweichung bei schiefen Verteilungen und extremen Beobachtungen. Ein pauschales Entfernen des jeweils kleinsten und größten Wertes ist nicht vorgesehen. Potenzielle Extremwerte werden anhand einer vorab festgelegten, für alle Items einheitlichen Regel identifiziert und anschließend anhand der Prozessdaten geprüft. Ein Ausschluss erfolgt nur bei begründetem Verdacht, dass der Wert die intendierte Itembearbeitungszeit nicht valide repräsentiert. Ausschlüsse werden dokumentiert. Bei relevanten Fällen werden Sensitivitätsanalysen mit und ohne ausgeschlossene Beobachtungen durchgeführt.

Sehr kurze Bearbeitungszeiten

Extrem kurze Bearbeitungszeiten bedürfen ebenfalls einer Prüfung. Eine schnelle Antwort kann durch sicheres Wissen oder eine leicht erkennbare Lösung entstehen. Auch ein oberflächliches Durchklicken kommt als Erklärung infrage. Qualitative Rückmeldungen aus der Pilotierung weisen darauf hin, dass Teilnehmer mitunter lediglich eine Auswahl treffen, um fortzufahren. Solche Rückmeldungen geben Anlass zur Prüfung des Antwortverhaltens. Dafür werden Kurzzeiten zusammen mit der Antwortkorrektheit und dem MC-Auswahlmuster betrachtet. Die Itemposition und die Zeiten benachbarter Items ergänzen die Einordnung.

Weiterentwicklung der Zeitmessung

Für die nächste technische Entwicklungsphase soll präziser zwischen Verweildauer und aktiver Bearbeitung unterschieden werden. Relevante Prozessdaten sind insbesondere Itemposition, Zeitpunkt der ersten Interaktion, Zeitpunkt der letzten Antwortänderung, Wechsel zum nächsten Item sowie – beim letzten Item – die Zeit zwischen letzter Interaktion und endgültiger Abgabe. Ein hartes Zeitlimit pro Item bleibt vorerst offen. Ein solches Limit würde Bearbeitungsgeschwindigkeit und Zeitdruck stärker in das gemessene Konstrukt einbeziehen und könnte die Validität beeinflussen. Ein mögliches Gesamtzeitfenster für spätere Prüfungs- oder Lernstandsszenarien soll erst auf Grundlage der empirischen Zeitdaten geprüft werden.

Methodische Leitlinie

Ergänzend wird die quantitative Analyse künftig systematisch mit einer qualitativen Itemprüfung verbunden. Empirische Schwierigkeit und AFB-Zuordnung werden getrennt beurteilt. Zusätzlich werden Itemstamm, fachliche und curriculare Passung, sprachliche Präzision sowie die Plausibilität der Distraktoren geprüft. PAL095 zeigt exemplarisch, dass ein statistisch brauchbares Item dennoch qualitativen Prüfbedarf aufweisen kann. Eine Revision erfordert eine begründete Entscheidung. Auch niederschwellige Items können im Gesamtblueprint eine sinnvolle Funktion erfüllen. Substanziell veränderte Items werden anschließend erneut pilotiert.

Die Pilotierung untersucht das gesamte Messsystem. Sie bezieht die Itemqualität und Auswahlarchitektur ein. Antwortverhalten und technische Prozessmessung gehören ebenfalls zur Prüfung. Auf eine statistische Auffälligkeit folgt die Analyse der Prozessdaten. Antwortmuster und Kontextinformationen ergänzen die fachliche Beurteilung. Erst danach wird über eine Revision oder einen begründeten Datenausschluss entschieden. Die bisherigen Daten bleiben erhalten und werden unter Berücksichtigung ihrer Erhebungsbedingungen ausgewertet.

Langfassung

1. Ziel und Datenbasis

Der PAL-Test V4.0 ist ein digitaler Lernstandstest in Entwicklung. Sein aktiver Pool umfasst 120 Items verschiedener Themenkategorien. Verwendet werden Single-Choice-Formate (SC) und Multiple-Choice-Formate (MC). Die Pilotierung dient der Gewinnung von Itemkennwerten und der Prüfung der technischen Testarchitektur.

Zum aktuellen Auswertungszeitpunkt lagen 99 registrierte Testdurchläufe vor. Vor der Analyse wurden Durchläufe ausgeschlossen, deren verwendetes Pseudonym die Zeichenfolge „test“ enthielt und die damit eigenen beziehungsweise technischen Erprobungen zugeordnet werden konnten. Nach Anwendung dieses vorab für die aktuelle Auswertung festgelegten Kriteriums verblieben 78 reguläre Durchläufe. Aus den Itemdaten lagen 1.350 auswertbare Itembearbeitungen vor. Alle 120 aktiven Items waren mindestens einmal exponiert.

2. Befund zur Itemexposition

Die Expositionshäufigkeit war deutlich ungleich verteilt. Pro Item lagen zwischen $n = 2$ und $n = 42$ Bearbeitungen vor. Der Median betrug 9,5. 108 der 120 Items erreichten mindestens fünf, 60 mindestens zehn, 33 mindestens 15 und 15 mindestens 20 Bearbeitungen. Der gesamte Pool ist damit zwar in der Pilotierung vertreten, die statistische Informationsbasis ist zwischen den Items jedoch noch deutlich unterschiedlich.

PAL095 stellt mit $n = 42$ das derzeit am häufigsten exponierte Item dar. PAL019 wurde bislang nur zweimal präsentiert. Solche Unterschiede sind für die Interpretation von Schwierigkeits- und Trennschärfekennwerten relevant, weil deren Stabilität wesentlich von der jeweiligen Beobachtungszahl abhängt.

3 Auswahl nach Themenkategorien und Antwortformaten

Die technische Prüfung der V4.0-Auswahlroutine zeigte, dass keine einfache Zufallsstichprobe aus dem gesamten Itempool gezogen wurde. Die Items werden zunächst nach Themenkategorien gruppiert. Die Kategorien werden in zufälliger Reihenfolge durchlaufen, während die Auswahlroutine zugleich versucht, ein annähernd ausgeglichenes Verhältnis von SC- und MC-Items herzustellen.

Diese Konstruktion ist für einen späteren Lernstandstest didaktisch nachvollziehbar: Gerade bei kurzen Durchläufen mit 10 oder 20 Fragen soll eine zufällige thematische Häufung vermieden und eine breite Inhaltsabdeckung gefördert werden. Für die psychometrische Pilotierung entsteht jedoch ein Zielkonflikt. Da die Kategorien unterschiedlich viele Items enthalten, verteilt sich eine annähernd gleiche Zahl von Kategorieziehungen auf unterschiedlich große Itempools. Items aus kleinen Kategorien erhalten dadurch im Mittel höhere Auswahlchancen als Items aus umfangreicheren Kategorien. Die zusätzliche SC-/MC-Steuerung kann diesen Effekt bei bestimmten Kategorie- und Typkonstellationen weiter verstärken.

Die Zufallskomponenten funktionieren innerhalb des implementierten Verfahrens. Die vorgeschaltete Balancing-Logik begünstigt jedoch systematische Unterschiede der Itemexposition. Für die Ziele der Pilotierung bedarf diese Auswahlarchitektur deshalb einer Überarbeitung.

4 Revision der Ziehungslogik bei stabilem Itempool

Die 120 Items bleiben zunächst inhaltlich unverändert. So lässt sich die Wirkung einer veränderten Auswahlarchitektur gesondert untersuchen. Für inhaltliche Revisionen muss zugleich eine ausreichend belastbare Datengrundlage entstehen.

Für die weitere Pilotierungsphase soll die Ziehungslogik expositionssensitiver gestaltet werden. Bislang selten präsentierte Items sollen eine höhere Auswahlwahrscheinlichkeit erhalten, während bereits stark exponierte Items vorübergehend geringer gewichtet werden. Inhaltliche Breite und eine angemessene Mischung von SC-

und MC-Items bleiben relevante Nebenbedingungen, sollen jedoch nicht erneut zu ausgeprägten systematischen Expositionsunterschieden führen.

Die technische Änderung wird mit Datum beziehungsweise eindeutigem Versions- oder Phasenbezug dokumentiert. Spätere Analysen müssen unterscheiden können, welche Antworten unter der bisherigen Auswahlroutine und welche unter der revidierten Routine entstanden sind. Die vor der Revision erhobenen Daten bleiben erhalten.

5. Vorgehen bei ersten Itemanalysen

Erste explorative Itemanalysen sind trotz der ungleichen Exposition möglich. Maßgeblich ist die Beobachtungszahl je Item. Sie ist von der Gesamtzahl der Testdurchläufe zu unterscheiden. Die Unsicherheit kleiner Stichproben muss bei der Interpretation berücksichtigt werden.

Als interne Arbeitsregel werden Items mit $n < 10$ überwiegend deskriptiv beobachtet. Bei $n = 10-19$ kann ein orientierendes Screening erfolgen. Ab etwa $n \geq 20$ ist eine erste vertiefte Zwischenanalyse vorgesehen. Diese Bereiche sind pragmatische Entscheidungshilfen. Eine Revision erfordert eine weitergehende Beurteilung des jeweiligen Items.

Die Itembeurteilung erfolgt mehrdimensional. Berücksichtigt werden insbesondere Schwierigkeitsindex beziehungsweise p-Wert, korrigierte Item-Gesamttest-Korrelation als Trennschärfemaß, SC-Distraktoren beziehungsweise MC-Auswahl- und Fehlermuster, Bearbeitungszeiten sowie – soweit sinnvoll – das subjektive Schwierigkeitsrating. Statistische Auffälligkeiten werden mit einer fachlich-inhaltlichen Prüfung des Items verbunden. Eine Itemlöschung oder inhaltliche Änderung erfolgt nicht allein aufgrund eines einzelnen vorläufigen Kennwerts.

6. Exemplarischer Befund PAL095

PAL095 ist mit $n = 42$ derzeit das am häufigsten bearbeitete Item und eignet sich deshalb als exemplarischer Fall für die Entwicklung der Auswertungslogik. Neben den klassischen Itemkennwerten fiel ein registrierter Zeitwert von rund 612,7 Sekunden auf.

PAL095 stand im betreffenden 20-Item-Durchlauf an erster Position. Die erste und letzte registrierte Interaktion lagen deutlich vor dem Ende des ausgewiesenen Zeitfensters. Dies deutet darauf hin, dass die gespeicherte Itemzeit auch Phasen außerhalb der aktiven Bearbeitung enthalten kann.

Der zweitlängste PAL095-Zeitwert von rund 105,0 Sekunden entstand bei Position 7 von 10. Auch hier erklärt eine verspätete Gesamtabgabe den Befund nur unzureichend. Vor einer abschließenden Interpretation muss die technische Definition der Zeitvariable geprüft werden.

7. Quantitative Beschreibung von Bearbeitungszeiten

Bearbeitungszeit ist eine metrische Zeitvariable. Arithmetisches Mittel und Standardabweichung sind daher grundsätzlich zulässige deskriptive Kennwerte. Ihre Aussagekraft hängt jedoch von der empirischen Verteilung ab. Bearbeitungszeiten können deutlich rechtsschief sein und einzelne sehr große Werte können Mittelwert und Standardabweichung stark beeinflussen.

Für die Pilotierung werden Mittelwert und Standardabweichung berichtet. Median und Quartile ergänzen die Beschreibung. Hinzu kommen Minimum und Maximum. Der Median liefert bei schiefen Verteilungen eine robuste Beschreibung der Lage. Der Interquartilsabstand (IQR) ergänzt die Standardabweichung als robustes Streuungsmaß.

Auch der kleinste und der größte Messwert bleiben zunächst im Datensatz. Ihr Ausschluss erfordert eine inhaltliche Begründung. Die bloße Rangposition unterscheidet valide Messwerte und Messartefakte nur unzureichend.

8 Prüfung auffälliger Bearbeitungszeiten

Auffällige Bearbeitungszeiten werden künftig in einem zweistufigen Verfahren behandelt. Zunächst erfolgt eine statistische Kennzeichnung potenzieller Extremwerte nach einer vorab festgelegten, für alle Items einheitlichen Regel. Als mögliche deskriptive Screeningregel kann beispielsweise die Tukey-Regel auf Basis des Interquartilsabstands verwendet werden. Die statistische Kennzeichnung führt jedoch nicht automatisch zum Ausschluss.

Im zweiten Schritt werden die vorhandenen Prozessdaten geprüft. Dazu gehören insbesondere Itemposition, erste und letzte registrierte Interaktion, Antwortänderungen, Wechsel zum nächsten Item, zeitliche Muster der unmittelbar benachbarten Items und gegebenenfalls die Zeit bis zur endgültigen Testabgabe.

Eine Beobachtung kann aus der Zeitanalyse ausgeschlossen werden, wenn ein begründeter Verdacht auf ein Messartefakt oder längere Inaktivität besteht. Entscheidend ist, ob der Wert die intendierte aktive Bearbeitung repräsentiert. Die Höhe oder Kürze des Wertes gibt zunächst nur Anlass zur Prüfung. Die angewandten Kriterien und die Anzahl der Ausschlüsse werden dokumentiert.

Bei relevanten Ausschlüssen erfolgt zusätzlich eine Sensitivitätsanalyse. Lage- und Streuungskennwerte werden für die vollständigen Daten und – sofern begründet – für den bereinigten Datensatz gegenübergestellt. Dadurch bleibt transparent, in welchem Ausmaß einzelne Extremwerte die Schlussfolgerungen beeinflussen.

9. Extrem kurze Zeiten und mögliches Rapid-Response-Verhalten

Sehr kurze Bearbeitungszeiten können durch eine unmittelbar sichere Lösung entstehen. Auch eine leicht erkennbare Antwortalternative kann die Bearbeitung verkürzen. Daneben kommt oberflächliches Durchklicken als Erklärung infrage. Dabei wird lediglich eine Auswahl getroffen, um zum nächsten Item oder zum Testabschluss zu gelangen.

Aus Gesprächen mit Teilnehmern liegen qualitative Hinweise auf solches Durchklicken vor. Sie werden als Begleitbeobachtungen behandelt. Eine Einordnung einzelner Durchläufe erfordert eine ergänzende Prüfung der jeweiligen Antwortmuster.

Bei auffällig kurzen Zeiten sollen daher zusätzliche Muster geprüft werden: Korrektheit der Antwort, Zahl ausgewählter MC-Alternativen, Itemposition, Bearbeitungszeiten benachbarter Items sowie Häufungen sehr kurzer Antworten innerhalb desselben Durchlaufs. Ein Muster aus mehreren extrem kurzen Bearbeitungen, auffälligen Antwortmustern und hoher Fehlerquote kann einen stärkeren Hinweis auf nicht ernsthaftes Response-Verhalten liefern als ein einzelner kurzer Messwert.

10. Präzisierung der technischen Zeitmessung

Die Pilotierungsdaten zeigen, dass zwischen reiner Verweildauer auf einem Item und aktiver Bearbeitung unterschieden werden sollte. Für die nächste technische Entwicklungsphase soll geprüft werden, welche Ereignisse die bisherige Variable `item_time_ms` tatsächlich umfasst und wie sie präziser operationalisiert werden kann.

Sinnvolle Prozessvariablen sind mindestens die Position des Items im Durchlauf, Zeitpunkt der ersten Interaktion, Zeitpunkt der letzten Antwortänderung, Zeitpunkt des Wechsels zum nächsten Item und – beim letzten Item – die Zeit zwischen letzter Interaktion und endgültiger Abgabe. Auf dieser Grundlage kann eine aktive beziehungsweise interaktionsnahe Bearbeitungszeit von bloßer Inaktivitäts- oder Verweildauer besser abgegrenzt werden.

Vor einer weitergehenden statistischen Interpretation der Zeitvariable ist diese technische Validierung prioritär.

11. Zeitlimits als mögliche spätere Testeigenschaft

Ein hartes Maximalzeitfenster pro Item bleibt vorerst offen. Seine Einführung würde das Prüfungsformat verändern. Die beobachtete Leistung wäre stärker von der Bearbeitungsgeschwindigkeit und vom Zeitdruck abhängig. Möglicher Prüfungsstress müsste ebenfalls berücksichtigt werden. Ob dies angemessen ist, hängt davon ab, welche Fähigkeit der Test erfassen soll.

Grundsätzlich besitzen auch Klausuren und Prüfungen ein festgelegtes Gesamtzeitfenster. Für einen späteren Einsatz des PAL-Tests kann deshalb ein Gesamtzeitlimit abhängig von der Zahl der ausgewählten Items geprüft werden. Eine solche Festlegung soll jedoch nicht willkürlich erfolgen, sondern erst nach Analyse der empirischen Bearbeitungszeitverteilungen und unter Berücksichtigung einer angemessenen Zeitreserve. Während der Pilotierung sind auch ungewöhnlich lange und kurze Zeiten zunächst wertvolle Prozessdaten.

12. Übergreifende Prüflogik und wissenschaftliche Redlichkeit

Aus den bisherigen Befunden wird eine allgemeine Blaupause für die weitere Pilotierung abgeleitet:

1. statistische Auffälligkeit identifizieren,
2. zugrunde liegende Beobachtungszahl und Verteilung prüfen,
3. Prozessdaten und Itemposition analysieren,
4. Antwort- beziehungsweise Distraktorenmuster einbeziehen,
5. qualitative Kontextinformationen nur als ergänzende Hinweise verwenden,
6. fachlich-inhaltliche Plausibilität prüfen,
7. erst anschließend über Beibehaltung, Beobachtung, Revision oder begründeten Datenausschluss entscheiden.

Die Pilotierung prüft das gesamte Messsystem. Dazu gehören der Itempool und die Ziehungslogik. Antwortverhalten und Prozessdaten werden im Zusammenhang mit der technischen Implementierung untersucht. Methodische Schwächen werden dokumentiert und begründet überarbeitet. Bisherige Daten bleiben erhalten. Änderungen werden zeitlich und anhand der jeweiligen Version gekennzeichnet, damit ihr Einfluss auf spätere Interpretationen nachvollziehbar bleibt.

13. Qualitative Itemrevision nach empirischer Analyse

Die quantitative Itemanalyse bildet nicht allein die Grundlage für Revisionsentscheidungen. Hinreichend exponierte oder statistisch auffällige Items werden zusätzlich qualitativ geprüft. Dabei sind insbesondere Itemstamm, fachliche und curriculare Passung, sprachliche Präzision, kognitive Anforderung, AFB-Zuordnung sowie die Qualität und Plausibilität der Antwortalternativen beziehungsweise Distraktoren zu beurteilen.

Empirische Itemschwierigkeit und Anforderungsbereich sind ausdrücklich voneinander zu unterscheiden. Der Schwierigkeitsindex beziehungsweise p-Wert beschreibt den Anteil korrekter Lösungen in der untersuchten Stichprobe. Der AFB beschreibt dagegen die intendierte kognitive Anforderung der Aufgabe. Ein empirisch leichtes Item ist daher nicht automatisch einem niedrigen AFB zuzuordnen. Umgekehrt begründet eine hohe formale AFB-Zuordnung keine hohe empirische Schwierigkeit. Beide Ebenen müssen getrennt beurteilt und anschließend gemeinsam interpretiert werden.

PAL095 verdeutlicht diese Unterscheidung exemplarisch. Das Item weist bei derzeit $n = 42$ eine Lösungsquote von rund 74 % und eine korrigierte Trennschärfe von etwa $r = 0,31$ auf und erscheint damit quantitativ zunächst brauchbar. Die qualitative Prüfung wirft jedoch die Frage auf, ob die bisherige Zuordnung zu AFB 3 die tatsächlich erforderliche kognitive Leistung angemessen beschreibt. Die Aufgabe verlangt vor allem das Erkennen copingfördernder Maßnahmen. Zudem erscheint die falsche Alternative semantisch vergleichsweise

eindeutig. Damit ist zu prüfen, ob das Item eher AFB 2 zuzuordnen ist und ob die Distraktorenqualität durch eine plausiblere, fachlich dennoch falsche Alternative verbessert werden könnte.

Aus dieser qualitativen Auffälligkeit folgt zum gegenwärtigen Zeitpunkt keine automatische Revision. Ein leichteres beziehungsweise niederschwelliges Item kann innerhalb eines 120-Item-Pools eine sinnvolle Funktion erfüllen, etwa zur Abdeckung basaler Kompetenzen und zur Gestaltung eines ausgewogenen Schwierigkeitsprofils. Entscheidend ist daher nicht die isolierte Bewertung eines Items, sondern seine Funktion im Testblueprint und im Zusammenspiel mit den übrigen Items.

Für künftige Revisionsentscheidungen gilt deshalb die folgende Reihenfolge: quantitative Analyse, qualitative Itemprüfung, Einordnung in den Testblueprint und anschließend die Entscheidung über Beibehaltung, Beobachtung, formale Umklassifizierung, inhaltliche Revision oder Entfernung. Substanziell veränderte Items sind psychometrisch als neue beziehungsweise neu zu pilotierende Items zu behandeln. Kennwerte vor und nach einer solchen Revision dürfen nicht unkritisch zusammengeführt werden.

14. Aktueller Entscheidungsstand

Bereich	Aktueller Befund	Konsequenz
Itempool	120 aktive Items. Stark unterschiedliche Exposition	Items zunächst unverändert lassen
Ziehungslogik	Kategorie- und SC/MC-Balancing erzeugt ungleiche Auswahlchancen	expositionssensitiv revidieren und Änderung dokumentieren
Itemanalyse	15 Items aktuell mit $n \geq 20$	explorative vertiefte Zwischenanalyse möglich
Zeitdaten	Extremwerte und Diskrepanz zwischen Verweildauer und Interaktionsdaten	Zeitmessung technisch validieren
Ausreißer	Extremwerte können Kennwerte bei kleinen n stark beeinflussen	standardisiert markieren, kontextuell prüfen, begründet exkludieren, Sensitivitätsanalyse
Sehr kurze Zeiten	möglicherweise Kompetenz, leichtes Item oder Rapid Response	Response Pattern statt Einzelwert prüfen
Zeitlimit	noch keine empirisch begründete Grenze	vorerst kein Itemlimit. Später Gesamtzeitfenster prüfen
Qualitative Itemrevision	Quantitative Kennwerte allein reichen nicht. PAL095 zeigt möglichen Prüfbedarf bei AFB und Distraktor	AFB, Itemstamm, Semantik, Distraktoren und Blueprint gemeinsam prüfen. Substanzielle Revision erneut pilotieren