children

MDPI

Systematic Review

Measurement Properties of Mallet Classification Systems in Individuals with Brachial Plexus Birth Injury: A COSMIN Systematic Review

Ali Ömer Acar [1,2,*], Barış Celbek [3] and Zeynep Hoşbay [4]

[1] Department of Physiotherapy and Rehabilitation, Faculty of Health Sciences, Acibadem Mehmet Ali Aydinlar University, 34752 Istanbul, Türkiye
[2] Department of Physiotherapy and Rehabilitation, Institute of Graduate Studies, Biruni University, 34015 Istanbul, Türkiye
[3] Physiotherapy Program, Department of Medical Services and Techniques, Vocational School of Health Services, Istanbul Aydin University, 34295 Istanbul, Türkiye; bariscelbek@aydin.edu.tr
[4] Department of Physiotherapy and Rehabilitation, Faculty of Health Sciences, Biruni University, 34015 Istanbul, Türkiye; zhosbay@biruni.edu.tr
* Correspondence: aliomer.acar@acibadem.edu.tr

Highlights

What are the main findings?

- "Mallet" is not a single instrument: seven review-defined operationalisations were identified across thirteen studies, and no included study evaluated an eligible measurement property of the original patient-level Mallet grading system.
- Task-level reliability varied across movements, while construct-validity findings were more consistent at the task level than at the composite level; certainty was low or very low, no eligible evidence on measurement error was identified, and major evidence gaps persisted.

What are the implications of the main findings?

- Mallet results should be reported explicitly by operationalisation, task set, and scoring level; the dimensionality and measurement properties of composite or total scores remain unestablished, and any preference for task-level reporting is provisional.
- Mallet grades should not be used as proxies for angle-defined glenohumeral motion; instrumented motion analysis is required for precise quantification, while Mallet systems remain practical complementary functional measures.

Abstract

Background/Objectives: The Mallet classification systems are widely used clinician-reported measures of shoulder function in children with brachial plexus birth injury (BPBI). However, "Mallet" encompasses multiple forms differing in task content, scoring, and interpretation. This review aimed to critically evaluate their measurement properties using COSMIN methodology. **Methods:** This review followed COSMIN 2.0, PRISMA 2020, and PRISMA-COSMIN for OMIs 2024 guidance and was registered in PROSPERO (CRD420261457308). Five databases were searched from inception to 23 July 2026, supplemented by grey literature and citation searches. Methodological quality was assessed using COSMIN Risk of Bias tools, and certainty was graded using the modified GRADE approach. **Results:** Thirteen studies met the eligibility criteria and represented seven review-defined Mallet operationalisations; no included study evaluated an eligible measurement property of the original patient-level Mallet grade. Task-level reliability varied by task, and task-level construct-validity findings more often met the review hypotheses than

Check for updates

Academic Editor: Carlo Camathias

Received: 28 July 2026
Revised: 22 October 2026
Accepted: 23 October 2026
Published: 18 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

composite-score findings; however, certainty was low or very low. Clinical grades generally overestimated angle-defined movement, particularly external rotation. In one low-certainty study, reliability varied more across rater configurations than across assessment modes. No eligible evidence was identified for measurement error, responsiveness, content validity, cross-cultural validity, or minimal important change, and structural-validity evidence was absent for multi-item forms used to generate composite scores. **Conclusions:** Mallet operationalisations should not be treated as interchangeable instruments. Task-level findings require cautious, operationalisation-specific interpretation, while the dimensionality and measurement error of composite scores remain unestablished; any preference for V1 task-level reporting is provisional.

Keywords: brachial plexus birth injury; Mallet classification; shoulder function; measurement properties; COSMIN; reliability; validity; systematic review

1. Introduction

Brachial plexus birth injury (BPBI) can cause persistent shoulder and upper-limb impairment that limits daily activities and participation in children who do not achieve full neurological recovery [1–3]. Because follow-up typically extends from infancy into adolescence, outcome assessment is needed to monitor recovery, guide surgical and rehabilitation decisions, and evaluate treatment [3,4]; however, heterogeneous assessment methods and reporting strategies have hindered comparison across studies and centres [5–7]. Mallet classification systems are among the most widely used task-based clinician-reported observational measures of shoulder function in children able to cooperate with the examination [4,8–11], whereas infants and non-cooperative children are usually assessed with measures such as the Active Movement Scale [12]. Their simplicity and lack of equipment requirements have supported wide clinical and research use [4,11], and the International Plexus Outcome Study Group (iPLUTO) consensus recommended a Mallet measure as a core outcome for children with BPBI [7].

Although frequently referred to under a single name, "Mallet" does not represent a single standardised measurement instrument. The original Mallet grading system assigns one patient-level ordinal grade from I to V based on the simultaneous performance of multiple shoulder movements. By contrast, the five-task item-scored modified Mallet independently grades global abduction, global external rotation, hand to neck, hand to spine, and hand to mouth from I to V and may sum the five item scores to produce a total score ranging from 5 to 25 [9]. Later forms added tasks such as hand to belly/internal rotation, supination, or an arm-at-rest classification and used varying item-level, summed, mean, selected-item, or overall-grade reporting structures. Because author terminology was inconsistent, these forms should be distinguished by their task content and scoring or reporting structure rather than by the labels "original", "modified", or "extended" alone. This distinction is clinically relevant because Al-Qattan and El-Sayed found that the original patient-level grade could not be assigned in 20% of 50 children when abduction and external rotation corresponded to different grade categories [13]. Furthermore, several Mallet tasks represent complex, multiplanar functional movements requiring coordinated contributions from the glenohumeral joint, scapulothoracic articulation, and trunk rather than isolated shoulder motion. For example, successful performance of the hand-to-spine task requires both glenohumeral internal rotation and shoulder extension, whereas many children who can complete the hand-to-belly task are nevertheless unable to reach behind the back, suggesting that these tasks evaluate different biomechanical constructs rather than the same aspect of shoulder function [14]. Similarly, recent kinematic studies have demonstrated

that compensatory trunk and scapular movements may substantially influence observed Mallet performance, meaning that observational scores may not always accurately reflect isolated glenohumeral motion [11,15,16]. Moreover, reported associations between Mallet scores and other BPBI outcome measures have varied, and measurement performance has been evaluated across face-to-face, telemedicine, and video-based assessments [8,9,17].

Evidence regarding the measurement properties of the Mallet classification systems has therefore accumulated across studies investigating reliability, validity, construct interpretation, biomechanical performance, and alternative modes of assessment. However, this evidence remains fragmented because studies have evaluated different task sets, scoring or reporting structures, patient populations, and assessment methods. Consequently, the available literature does not provide a clear understanding of the measurement properties of each Mallet operationalisation or whether findings obtained from one task set or scoring structure can be generalised to another. COSMIN provides a standardised methodology for critically appraising study quality and synthesising evidence on the measurement properties—including reliability, validity, and responsiveness of health outcome instruments [18]. Previous reviews have summarised outcome assessment methods for brachial plexus birth injury more broadly [6], but to our knowledge, no prior review has applied the COSMIN methodology specifically to the Mallet classification systems while explicitly distinguishing between different versions rather than treating "Mallet" as a single homogeneous instrument. An operationalisation-specific COSMIN-based systematic review is therefore needed to critically appraise the methodological quality of existing studies, summarise the evidence for each measurement property, and determine the certainty of the available evidence.

This review therefore aimed to use COSMIN methodology to appraise the methodological quality of the included studies, synthesise the evidence on each measurement property, and grade its certainty separately for each Mallet operationalisation, task, and scoring level (individual task versus composite score).

2. Materials and Methods

2.1. Design and Reporting Framework

This instrument-focused systematic review evaluated the measurement properties of Mallet classification systems used in brachial plexus birth injury. COSMIN version 2.0 was the primary methodological reference [18], supplemented where relevant by the earlier COSMIN systematic-review guidance [19]. Reporting followed PRISMA 2020 [20] and PRISMA-COSMIN for OMIs 2024 [21], and measurement-property terminology followed the COSMIN taxonomy [22].

Because Mallet classifications are clinician-reported observational measures (ClinROMs), COSMIN guidance developed primarily for patient-reported outcome measures was adapted to clinician-observed task content, scoring instructions, and rater-based ratings. Content-validity concepts were interpreted in relation to the relevance, comprehensiveness, and comprehensibility of the task set and scoring system for the intended clinical construct and population. For reliability and measurement error, the COSMIN quantitative criteria were retained: an appropriate reliability coefficient of at least 0.70 was considered sufficient, and measurement error was considered sufficient only when the smallest detectable change or limits of agreement were smaller than an established minimal important change. Construct-validity hypotheses and expected association magnitudes were review-defined in accordance with COSMIN hypothesis-testing principles and are specified in Section 2.6. Risk of bias for reliability and measurement error was assessed with the COSMIN Risk of Bias tool, which was developed for outcome measurement instruments of any type, including clinician-rated and performance-based instruments [23]; for hypothesis testing, the generic

COSMIN standards on comparator description, comparator measurement properties, and statistical methods were applied without modification [24]. PROM-specific standards on patient comprehensibility and translation were not applicable because no included study assessed content validity or cross-cultural validity. These adaptations were specified before data extraction.

The protocol was prospectively registered in PROSPERO before database searching and study screening (CRD420261457308). No eligibility criterion was changed after registration. Embase.com was subsequently added as an additional information source to strengthen database coverage, and the PROSPERO record and protocol were updated accordingly. Rating criteria and construct-validity hypotheses that were not specified in the registration were defined by the review team after data extraction, when the extracted results were known, but before formal COSMIN sufficiency rating and evidence synthesis; they are therefore reported as post hoc (Section 2.6).

2.2. Eligibility Criteria

Studies were eligible if participants had a brachial plexus injury sustained at birth, regardless of terminology, age, or assessment setting. Mixed-population studies required separately extractable BPBI data. Acquired traumatic, iatrogenic, neoplastic, and other non-birth-related injuries were excluded.

Eligible forms included: the original patient-level Mallet grading system [10], the five-task item-scored modified Mallet; individual task scores and derived summed, mean, or composite scores; six-task forms adding hand to belly/internal rotation; forms additionally incorporating supination or an arm-at-rest classification; and partial or unclearly derived Mallet uses. An operationalisation was defined as a distinct combination of task content and scoring or reporting structure. Partial use was classified separately when a study administered or operationally modified a specific subset or task rather than merely reporting selected results from a complete form. V1–V7 operationalisation assignments were made jointly by two reviewers through discussion and consensus during data extraction, rather than independently in duplicate, and were finalised before risk-of-bias assessment and evidence synthesis. A form was coded as a separate operationalisation when the administered task set differed (addition, omission, or substitution of a task, or administration of a single task only) or when item-level scores were unavailable because only an overall grade of undescribed derivation was reported. Different reporting levels of the same administered form (task scores, sum, mean, or selected items) were not coded separately but were stratified as scoring levels during synthesis; author labels or task wording alone were not considered sufficient. Author terminology was preserved during data extraction but was not used to assume measurement equivalence. The study-level mapping supporting these assignments is provided in Table S1.

Published original studies were included when an aim or extractable analysis addressed instrument development or a measurement property. Studies were excluded when Mallet scores served only as clinical, surgical, or treatment outcomes; when a Mallet task was used only as a movement prompt; or when no Mallet-specific measurement-property result was extractable. Studies focused primarily on another construct remained eligible when they contained an extractable Mallet-specific measurement-property analysis. Non-empirical publications, conference abstracts without full articles, and non-English full-text reports were excluded.

Convergent and known-group validity were evaluated under hypothesis testing in accordance with COSMIN guidance [18,19]. Reliability and measurement-error findings were classified according to the statistics reported. Kinematic analyses were classified according to their analytic target. Associations between Mallet task scores and related but distinct segmental or multi-joint kinematic constructs were evaluated as construct validity, whereas direct comparisons between a clinically assigned Mallet grade and a grade derived from angle-defined motion for the same task were evaluated as agreement with an angle-defined reference. Three-dimensional motion analysis was not treated as a gold standard for overall upper-limb function. Interpretability and feasibility were summarised descriptively and did not constitute measurement properties under the COSMIN taxonomy [18,22]. Evidence on these characteristics was extracted only from studies otherwise eligible on the basis of an instrument-development or measurement-property analysis.

2.3. Information Sources and Search Strategy

PubMed/MEDLINE, Scopus, CINAHL Ultimate (EBSCOhost), and Web of Science Core Collection were searched from inception to 23 July 2026 without year restrictions. No language filter was applied during the database searches; however, non-English reports were excluded during full-text eligibility assessment. Searches combined Mallet terms, controlled vocabulary and free-text BPBI synonyms, and measurement-property terms. The sensitive PubMed measurement-properties filter developed by Terwee et al. [25] was adapted for other databases. Embase (Embase.com) was additionally searched on 7 August 2026 using the same concept blocks translated to Emtree syntax, with a publication-date limit of 23 July 2026 to match the original search window; records were deduplicated against the existing set and screened using identical criteria.

Supplementary searches covered the first 100 Google Scholar results, iPLUTO and Pediatric Orthopaedic Society of North America outputs, reference lists, and forward citations. Records were deduplicated against the database set and screened using identical criteria. Full strategies are provided in Table S2.

2.4. Study Selection and Data Extraction

Two reviewers independently screened titles, abstracts, and full texts. One reviewer performed data extraction and COSMIN assessments, and a second reviewer verified all entries and judgements against the source and predefined decision rules. Disagreements were resolved by consensus or a third reviewer. Screening decisions and exclusion reasons for all records assessed at title/abstract and full-text level are provided in Table S3.

Records were managed and deduplicated in Rayyan (Rayyan Systems Inc., Cambridge, MA, USA) [26]. Titles and abstracts were screened hierarchically for population, instrument, analysis, and publication type; uncertain records underwent full-text review. Inclusion required confirmation of the BPBI population, an eligible Mallet system, and an extractable measurement-property analysis. One primary exclusion reason was recorded per excluded report, and potentially overlapping cohorts were linked to avoid double counting.

A piloted spreadsheet was used to extract study and participant characteristics, details of each Mallet version, assessment procedures, and measurement-property results. For each analysis, we recorded the sample, comparator, hypothesis, statistical method, estimate, confidence interval, and *p*-value where available. Reliability analyses were additionally documented by coefficient type, model, weighting, and rater design, while relevant error indices were recorded for measurement error. Values were reconstructed only when the required data and calculation were transparent and were labelled as reviewer-derived. Missing statistics were not imputed, and correlations were not converted into agreement coefficients.

2.5. Risk of Bias and Reporting Bias

Risk of bias was assessed separately for each measurement property and, where necessary, for individual analyses. The relevant COSMIN Risk of Bias tools were used for validity, reliability, and measurement error [18,23,24]. Relevant sources of variation were defined before the standards were applied. Each analysis was rated as very good, adequate, doubtful, or inadequate according to the COSMIN worst-score-counts principle.

Missing-result bias was assessed qualitatively by comparing methods and results and recording omitted or selectively reported coefficients, confidence intervals, task-level findings, or planned analyses. Statistical small-study-effect tests were not feasible.

2.6. Rating of Individual Results

Results were rated as sufficient, insufficient, or indeterminate using the updated COSMIN criteria [18]; interpretability and feasibility were summarised descriptively. Reliability was considered sufficient when an appropriate agreement coefficient—an intraclass correlation coefficient, generalisability coefficient, or weighted kappa—was at least 0.70. Pearson or Spearman correlations of repeated scores were considered in the risk-of-bias assessment, in which evidence of systematic differences between ratings was taken into account [23]; because they do not meet the COSMIN criterion for sufficient reliability, such results were rated as indeterminate [18]. Measurement error could be rated as sufficient only when the smallest detectable change or limits of agreement were below an established minimal important change.

Composite Mallet scores are used clinically as global summaries of shoulder function, but neither the original nor the modified systems specify whether the total reflects a single latent construct (a reflective model) or is defined by the combination of distinct task achievements (a formative model). Because the tasks assess biomechanically distinct movements [9,14] and no included study tested dimensionality, we did not assume a reflective model for composite scores. Under COSMIN, structural validity and internal consistency are relevant only to reflective models [18,19]. Cronbach's alpha was therefore not interpreted, and the absence of factor-analytic or Rasch evidence was reported as an evidence gap that applies only if composite scores are intended to represent a single latent construct. If composite scores are regarded as formative indices, structural validity and internal consistency are not applicable, and the principal unaddressed property is content validity, that is, whether the task set adequately represents the intended construct. Individual task grades were evaluated as single-item scores.

Construct validity was evaluated using review-defined hypotheses formulated after data extraction but before formal COSMIN rating and evidence synthesis. Because the extracted results were known at that stage, all review-defined hypotheses are reported as post hoc. To limit outcome-driven specification, each comparison was assigned to a hypothesis class on the basis of the conceptual and biomechanical relationship between the Mallet variable and the comparator, and correlation thresholds were adopted from COSMIN guidance rather than derived from the observed data [18,19]. Correlations $\geq$0.50 were expected for the same or closely related movement, $\geq$0.30 for related but non-identical constructs, and $|r|$ or $|\rho| < 0.30$ for anatomically distal or non-contributing relationships. Known-group hypotheses addressed injury extent and Narakas type. Circular, insufficiently defined, or unstandardised comparisons were not rated (Table S4).

Criterion validity required a defensible reference standard. Three-dimensional motion analysis was treated only as an angle-defined reference for the movement assessed, not as a gold standard for overall upper-limb function. A correlation or area under the curve of at least 0.70 was considered sufficient. For categorical task grades, kappa was the principal agreement statistic and, for dichotomous classification, both sensitivity and specificity had

to reach this threshold; exact agreement and rank correlations were supporting estimates. Because COSMIN provides no criteria for this comparison, these thresholds were review-specific. Results without an interpretable threshold were rated as indeterminate.

2.7. Evidence Synthesis and Certainty of the Evidence

Evidence was synthesised narratively within clinically comparable operationalisations, tasks, scoring levels, rater designs, and assessment modes. Meta-analysis was planned only when at least two studies were sufficiently comparable in population, assessment design, statistical method, and reported outcome; these conditions were not met. When multiple rated results contributed to a synthesis, ≥75% agreement was required for an overall sufficient or insufficient rating; otherwise, evidence was considered inconsistent. Indeterminate and not-rated results were excluded from this denominator. Repeated task, rater, mode, or encounter estimates were not treated as independent samples, and potentially overlapping cohorts were represented by the largest clearly unique sample. Result- and synthesis-level details, including downgrade reasons and unique participant counts, are provided in Tables S5 and S6.

Certainty was graded high, moderate, low, or very low using the COSMIN modified GRADE approach [18,19]. Evidence began at high certainty and was downgraded for risk of bias, inconsistency, imprecision, and indirectness. For single-study syntheses, risk of bias was downgraded one level for adequate, two for doubtful, and three for inadequate evidence; when several studies contributed, the COSMIN rules for the number and quality of studies were applied [18,19]. Inconsistent (±) syntheses were downgraded one level. Samples of at least 100 unique participants were not downgraded for imprecision, 50–99 were downgraded one level, and fewer than 50 were downgraded two levels. Certainty was graded separately when combining operationalisations, tasks, scoring or reporting levels, or assessment modes would obscure clinically relevant differences. Indeterminate results, evidence gaps, interpretability, and feasibility were not graded; publication bias was assessed separately.

3. Results

3.1. Study Selection and Study Characteristics

Thirteen studies published between 2003 and 2025 met the eligibility criteria (Figure 1). Analysed samples ranged from 15 to 193 participants, and reported source cohorts included participants aged 1 month to 19 years (Table 1). The lower bound derives from Bae et al. [8], who scored the modified Mallet in infants by eliciting movement with toys and parental assistance; agreement for some movements was relatively poor in children younger than two years, who were the least cooperative. This lower age limit should therefore not be interpreted as evidence that Mallet scoring is reliable or feasible in young infants. Sample sizes were not summed, because Greenhill et al. [9] contributed repeated encounters rather than independent participants, and the Shriners motion-capture reports [14–16] and the Hacettepe reports [17,27] may draw on overlapping cohorts.

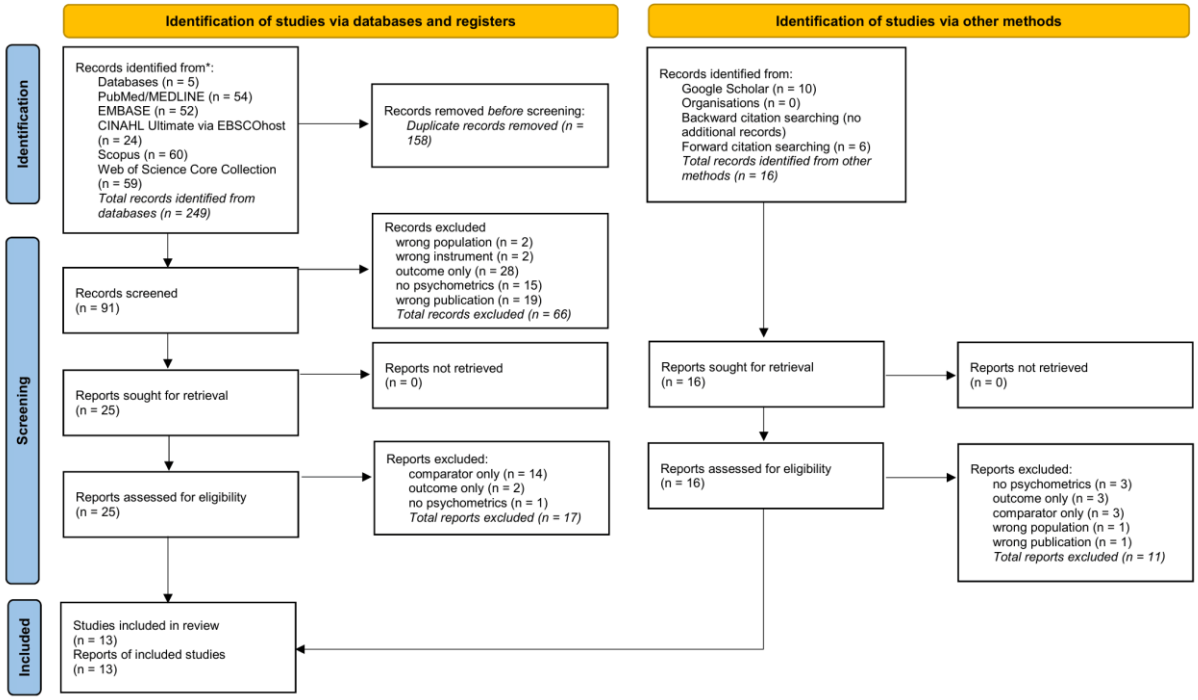


Figure 1. PRISMA 2020 Flow Diagram of Study Selection. * Records retrieved from each database. PubMed/MEDLINE, CINAHL Ultimate, Scopus, and Web of Science Core Collection were searched on 23 July 2026; Embase was searched on 7 August 2026 with a publication-date limit of 23 July 2026.

The evidence was unevenly distributed across measurement properties. Three studies examined reliability [8,17,28], of which only Bae et al. [8] also reported Cronbach's alpha, which was not interpreted. Eight contributed construct-validity evidence, and two compared clinical grades with angle-defined three-dimensional motion analysis [15,16]. No study evaluated responsiveness, content validity, or cross-cultural validity, and no multi-item composite form was evaluated for structural validity.

3.2. Mallet Operationalisations Identified

Seven review-defined Mallet operationalisations were identified (V1–V7; Table 2). V1 was the five-task item-scored modified Mallet [8,28,29]; V2 was a non-standard five-task form in which flexion replaced global external rotation [30]; V3 was the six-task item-scored extended form adding hand-to-belly/internal rotation [9,14–16,31]; V4 and V5 incorporated an arm-at-rest classification alongside seven- and six-task sets, respectively [17,27]; V6 represented unclear reporting of the six-task extended form as a single overall grade [32]; and V7 represented partial use of the global external-rotation item in shoulder adduction [33]. The original patient-level Mallet grade [10] was treated as a reference form (R0) and was not counted among the seven operationalisations because no included study evaluated an eligible measurement property of that form. The inability to assign an original Mallet grade in a subset of patients reported by Al-Qattan et al. [13] was considered an applicability/interpretability observation rather than an eligible measurement-property analysis. The task composition and scoring architecture of R0 and V1–V7 are compared in Figure 2.

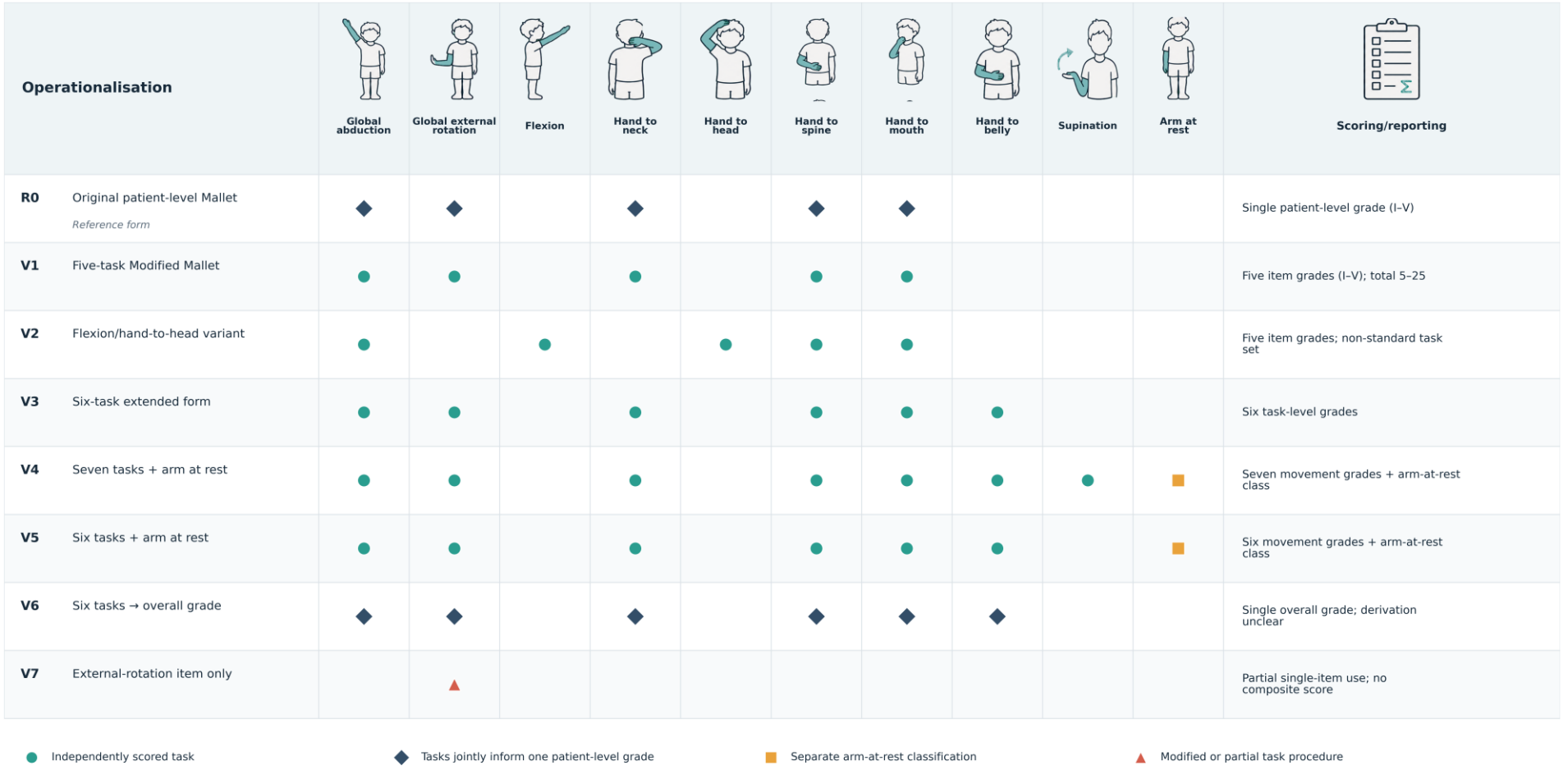


Figure 2. Structural comparison of the original Mallet reference form and seven review-defined operationalisations.

https://doi.org/10.3390/children13111604

Table 1. Characteristics of the included studies.

Study; Country	Sample and Age	Mallet Operationalisation	Evidence Contributed
Bae et al., 2003 [8]; USA	$N = 80$; mean 3.3 y (1 mo.–15 y)	V1: Five-task item-scored modified Mallet; sum 5–25	Reliability (Cronbach's α reported, not interpreted)
van der Sluijs et al., 2006 [28]; Netherlands	$N = 30$; mean 7.1 y (4.5–10 y)	V1: Five-task item-scored modified Mallet; tasks analysed separately	Inter-rater reliability
Bae et al., 2008 [29]; USA	$N = 150$; Mallet–PODCI $n = 76$; 2–5 y	V1: Five-task item-scored modified Mallet; sum 5–25	Construct validity; interpretability
Talbert et al., 2011 [33]; USA	$N = 74$; MRI subset $n = 28$; mean 5.1 y	V7: Partial single-item use; global external rotation assessed in shoulder adduction, grades I–V and dichotomised as ≤II versus ≥III	Construct validity
Brochard et al., 2014 [30]; USA–France	$n = 15$; mean 11.5 y (5–18 y)	V2: Non-standard five-task form; flexion, abduction, hand to head, hand to mouth, and hand to spine; sum 5–25	Construct validity
Ismaeel et al., 2014 [32]; Egypt	$N = 30$; mean 4.7 y (3.6–6 y)	V6: Six-task extended form reported as one overall grade (2–4); derivation not stated	Construct validity
Greenhill et al., 2017 [9]; USA	144 participants; repeated encounters; 2.1–11.4 y	V3: Six-task item-scored extended form; sum 6–30 and post hoc five-task total	Construct validity; interpretability; feasibility
Russo et al., 2019 [14]; USA	$N = 32$; mean 8.9 y (3–17 y)	V3: Six-task item-scored extended form; hand-to-spine and hand-to-belly analysed	Construct validity
Russo, Chafetz et al., 2022 [16]; USA	$N = 26$; mean 9.9 y (5–17 y)	V3: Six-task item-scored extended form; external-rotation item analysed	Agreement with motion analysis; interpretability
Russo, Topley et al., 2022 [31]; USA	$N = 80$; mean 10.1 y	V3: Six-task item-scored extended form; mean item score 1–5	Construct validity
Delioğlu et al., 2024 [17]; Türkiye	$N = 100$; mean 7.0 y (4–15 y)	V4: Seven-task extended form plus arm-at-rest classification; eight components analysed separately; no total	Reliability; feasibility
Lovette et al., 2024 [15]; USA	$N = 107$; mean 8.7 y (3.3–19 y)	V3: Six-task item-scored extended form; three items analysed	Agreement with angle-defined motion analysis; interpretability
Delioğlu et al., 2025 [27]; Türkiye	$N = 193$; mean 7.4 y (4–18 y)	V5: Six-task extended form plus arm-at-rest classification; individual tasks, resting grade, forearm category, and reported total-function score	Construct validity; interpretability

Note. PODCI = Paediatric Outcomes Data Collection Instrument. Sample sizes are not summed because of repeated encounters and possible participant overlap within the Shriners and Hacettepe report groups. V1–V7 indicate the review-defined operationalisations described in Table 2.

Table 2. Reference form and review-defined Mallet operationalisations.

Code	Operationalisation	Studies	Task Content	Scoring/Reporting Structure
R0	Original patient-level Mallet grade	No eligible included study; historical reference form	Five movements represented simultaneously within one patient-level grade	One ordinal grade, I–V; no independent item profile or summed score
V1	Five-task item-scored modified Mallet	Bae 2003 [8], 2008 [29]; van der Sluijs 2006 [28]	Global abduction, global external rotation, hand to neck, hand to spine, hand to mouth	Independent I–V task grades; tasks and/or sum 5–25
V2	Non-standard five-task item-scored Mallet form	Brochard 2014 [30]	Flexion, abduction, hand to head, hand to mouth, hand to spine	Independent scores 1–5; sum 5–25 and task scores
V3	Six-task item-scored extended Mallet form	Greenhill 2017 [9]; Russo 2019 [14]; both Russo 2022 reports [16,31]; Lovette 2024 [15]	V1 task set plus hand to belly/internal rotation	Independent I–V grades; sum 6–30, mean 1–5, post hoc five-task total, or selected items
V4	Seven-task extended form with arm-at-rest classification	Delioğlu 2024 [17]	V3 task set plus supination; separate arm-at-rest classification	Seven ordinal movement grades plus one arm-at-rest classification; eight components analysed separately; no total
V5	Six-task extended form with arm-at-rest classification	Delioğlu 2025 [27]	V3 task set plus separate arm-at-rest classification	Six movement grades; arm-at-rest grades 1–6 with pronation/supination subcategories; reported total-function score with calculation not described
V6	Unclear overall-grade reporting of a six-task extended form	Ismaeel 2014 [32]	Six-task modified Mallet illustration	One overall grade 2–4 reported per participant; derivation not stated
V7	Partial single-item modified Mallet use	Talbert 2011 [33]	Global external rotation only, assessed in shoulder adduction	Consensus grade I–V; analysed continuously and dichotomised as ≤II versus ≥III

Note. R0 denotes the original patient-level Mallet grade and is presented as a historical reference form; it is not counted among the seven operationalisations identified in the included studies. V1–V7 are review-defined labels based on task content and scoring or reporting structure. Author terminology, including "original", "modified", and "extended", was not assumed to indicate measurement equivalence. No included study evaluated an eligible measurement property of R0.

3.3. Risk of Bias

Methodological quality varied by analysis type (Table 3). Risk of bias was adequate for the task-specific weighted-kappa analyses of van der Sluijs et al. [28] and for the task-aggregated weighted-kappa analyses of Bae et al. [8], although Bae et al. reported means across tasks rather than task-specific estimates. For the aggregate score of Bae et al., reliability was adequate, whereas measurement error could not be assessed because no error index was reported. Delioğlu et al. [17] was rated doubtful because blinding was incompletely

described and administration mode was partly confounded with assessment occasion. Construct-validity analyses were predominantly doubtful because comparator measurement properties were unestablished, repeated encounters were not statistically accounted for, or only selected results were reported. Analyses were inadequate when the Mallet variable was insufficiently defined or when comparison groups were defined using the Mallet scores themselves.

Table 3. COSMIN risk-of-bias ratings by study and analysis.

Study/Analysis	Reliability	Measurement Error	Internal Consistency	Construct Validity	Agreement with Angle-Defined Motion Analysis
Bae et al., 2003 [8]: task ratings	A	—	—	—	—
Bae et al., 2003 [8]: aggregate score	A	—	NA	—	—
van der Sluijs et al., 2006 [28]	A	—	—	—	—
Bae et al., 2008 [29]	—	—	—	D	—
Talbert et al., 2011 [33]	—	—	—	D	—
Brochard et al., 2014 [30]	—	—	—	D	—
Ismaeel et al., 2014 [32]	—	—	—	I	—
Greenhill et al., 2017 [9]	—	—	—	D	—
Russo et al., 2019 [14]	—	—	—	D (correlations); I (score-defined groups)	—
Russo, Chafetz et al., 2022 [16]	—	—	—	—	I
Russo, Topley et al., 2022 [31]	—	—	—	D (convergent); A (injury groups); D (surgery groups)	—
Delioğlu et al., 2024 [17]	D	—	—	—	—
Lovette et al., 2024 [15]	—	—	—	—	D (abduction, external rotation); I (hand to mouth)
Delioğlu et al., 2025 [27]	—	—	—	D	—

Note. A = adequate; D = doubtful; I = inadequate; NA = not applicable; — = not evaluated or no eligible analysis. Ratings apply to individual analyses using the worst-score-counts principle. Motion analysis served as an angle-defined reference comparator and was not treated as a gold standard for overall function. Internal consistency was not applicable because a reflective model was not assumed for composite scores; no measurement-error index was reported.

3.4. Reliability and Measurement Error

Within V1, reliability was task-dependent rather than uniformly strong across the five-task set. In van der Sluijs et al. [28], hand-to-neck reliability was consistently sufficient and hand-to-mouth reliability was consistently insufficient. Global abduction, global external rotation, and hand-to-spine crossed the 0.70 threshold across observer pairs and were therefore rated inconsistent. Bae et al. [8] reported mean weighted kappas of 0.76 within and 0.78 between raters, but only means across tasks were available; these values could not resolve the task-specific inconsistencies. Certainty was very low for all five task-level syntheses.

Aggregate-score reliability was reported only as Pearson correlations of repeated totals (mean r = 0.92 within and r = 0.78 between raters; Bae et al. [8]). Because COSMIN requires an intraclass correlation coefficient or weighted kappa of at least 0.70 for sufficient reliability, these results were rated indeterminate and were not graded. No study reported a standard error of measurement, smallest detectable change, or limits of agreement, so no eligible evidence on measurement error was available. Cronbach's alpha was not interpreted because a reflective model was not assumed for composite scores and their dimensionality had not been established [8].

In the single V4 study, reliability varied more across rater configurations than across administration modes. Delioğlu et al. [17] found all within-rater cross-mode and two-rater estimates sufficient across face-to-face, live tele-, and recorded-video assessments ($\kappa = 0.79–1.00$). In four-rater remote assessments, reliability was sufficient for only global abduction, hand to spine, and hand to belly and insufficient for the other five components ($\kappa = 0.63–0.65$). Certainty was low for all of these syntheses.

3.5. Construct Validity

Movement- and score-specific comparisons generally provided clearer construct-validity evidence than composite-score comparisons. Eight studies contributed 63 construct-validity results, of which 27 were formally rated against the review hypotheses, and 36 were not rated. Unrated results included circular or same-framework comparisons [14,27], an insufficiently defined Mallet variable [32], unstandardised estimates [29], a confounded surgical-group comparison [31], and supporting analyses from the same comparator and sample that were not counted separately [33]. Result-level decisions are detailed in Table S5.

Task/score-level convergence was stronger than composite-level convergence. All six Mallet–Active Movement Scale task/score comparisons met expectations [9,12]. At the composite-score level, two of three comparisons met expectations. Relationships with the Active Movement Scale and the Brachial Plexus Profile Activity Short Form were sufficient [9,31], whereas the relationship with the Toronto Test Score was insufficient [9,34]. Six of eight task–kinematic correlations [14] and all five correlations between shoulder strength and the V2 non-standard task scores met expectations [30]. Evidence for muscle-related comparators was inconsistent: infraspinatus volume met expectations, but electromyographic interference grade did not [33].

Known-group findings were generally consistent. Better Mallet scores were observed in children with less extensive root-level injuries [31]. In the V5 study by Delioğlu et al. [27], arm-at-rest grades and associated forearm pronation/supination categories differed across Narakas injury types [35]. Secondary-surgery group differences were not rated because surgical indication was confounded with severity and age. Certainty was low for four syntheses and very low for the remaining rated syntheses, reflecting doubtful methodological quality, small contributing samples, and selective reporting of significant correlations.

3.6. Agreement with Angle-Defined Motion Analysis

Clinical Mallet grades showed insufficient agreement with motion-capture-derived grades for every item examined and generally overestimated angle-defined movement [15,16]. Russo et al. [16] found exact external-rotation agreement in 6 of 26 participants, with all disagreements favouring the clinical grade. Lovette et al. [15] reported near-zero agreement for external rotation ($\kappa = 0.065$), moderate agreement for global abduction ($\kappa = 0.50$), and low hand-to-mouth classification accuracy despite 80.0% sensitivity. Mean clinical overgrading was 1.03 grades for external rotation and 0.17 for abduction. Certainty was low for external rotation and abduction and very low for hand-to-mouth. Motion capture was treated only as an angle-defined reference for the movement assessed. Figure 3 summarises the direction and certainty of evidence across all graded reliability, construct-validity, and angle-defined agreement syntheses.

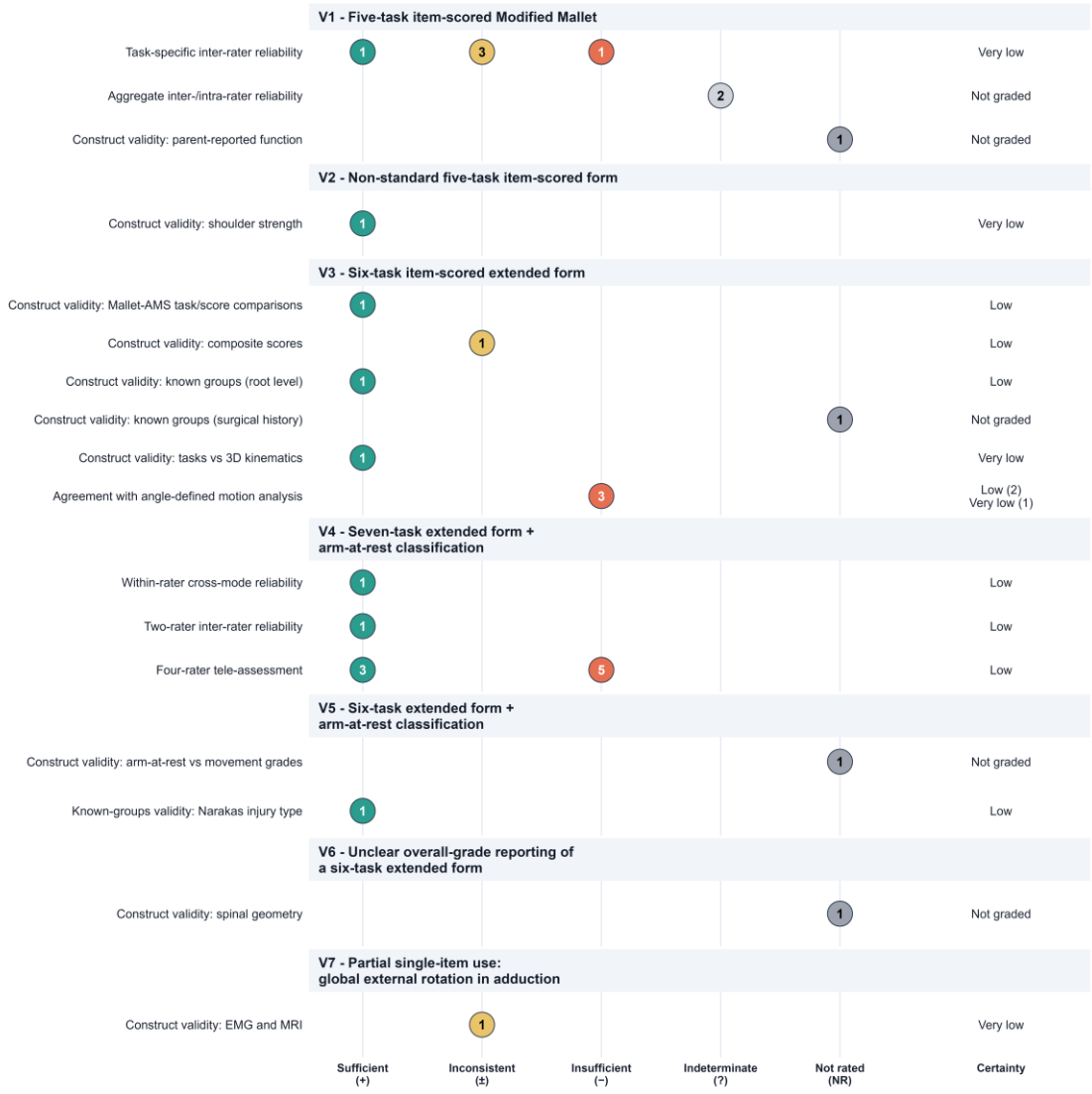


Figure 3. Direction and certainty of evidence for Mallet measurement properties. Numbers within circles indicate the number of task-, component-, or synthesis-level findings assigned to each rating within a row, rather than the number of participants or independent studies. + = sufficient; ± = inconsistent; − = insufficient; ? = indeterminate; NR = not rated. Certainty was graded only for formally rated findings. Operationalisation codes V1–V7 correspond to the forms described in Table 2. Internal consistency is not shown because a reflective model was not assumed for composite scores.

3.7. Interpretability, Feasibility and Evidence Gaps

Interpretability evidence was descriptive and did not establish severity thresholds, minimal important change, or item equivalence. Reported findings included associations between V1 total-score bands and parent-reported function [29], variable item–composite correlations [9], and age-related differences in resting posture among children with total-plexus injury [27]. Feasibility was limited by age, cooperation, contracture, connectivity, and caregiver participation; remote assessment took approximately twice as long as face-to-face assessment [9,17].

No eligible evidence was identified for responsiveness, content validity, or cross-cultural validity, and no study reported a minimal important change (Table 4). Structural-validity evidence, which would be required if composite scores were intended to reflect a single latent construct, was absent for multi-item forms used to generate composite scores; structural validity was not applicable to partial single-item or single-overall-grade uses. Meta-analysis was not appropriate because the studies differed substantially in Mallet operationalisation, task set, scoring or reporting level, rater design, assessment mode, comparator, and statistical method. Findings were therefore synthesised separately by version, task, rater, and setting.

Table 4. Summary of findings and certainty of evidence.

Op.	Property	Finding	Rating	Certainty
R0	All eligible measurement properties	No eligible evidence	NE	Not graded
V1	Task-level IRR	HN sufficient; HM insufficient; GA, GER, and HS inconsistent	+/−/±	Very low
V1	Aggregate inter-/intra-rater reliability	Pearson r only (0.92/0.78); no ICC or weighted κ	?	Not graded
V1	IC	α reported; not applicable because a reflective model was not assumed	NA	Not graded
V1	CV: parent-reported function	Unstandardised regression slopes only	NR	Not graded
V2	CV: shoulder strength	5/5 hypotheses met	+	Very low
V3	CV: task level–AMS task/score comparisons	6/6 hypotheses met	+	Low
V3	CV: composite vs. clinical scales	2/3 hypotheses met	±	Low
V3	CV: known groups—root level	Higher scores in less extensive injuries	+	Low
V3	CV: known groups—surgical history	Confounded by severity and age	NR	Not graded
V3	CV: tasks vs. 3D kinematics	6/8 hypotheses met	+	Very low
V3	Angle-defined agreement: ER	Near-zero ordinal agreement; systematic overgrading	−	Low
V3	Angle-defined agreement: GA	$\kappa = 0.50$ despite 75.7% exact agreement	−	Low

Table 4. *Cont.*

Op.	Property	Finding	Rating	Certainty
V3	Angle-defined agreement: HM	Low specificity and overall accuracy	−	Very low
V4	Cross-mode/two-rater reliability	All estimates sufficient (κ = 0.79–1.00)	+	Low
V4	Four-rater tele-reliability	3/8 components sufficient	+/−	Low
V5	CV: posture vs. movement grades	Same-framework comparator	NR	Not graded
V5	CV: known groups—Narakas injury type	2/2 hypotheses met	+	Low
V6	CV: spinal geometry	Analysed grade undefined	NR	Not graded
V7	CV: EMG and MRI	1/2 hypotheses met	±	Very low
All	Measurement error; responsiveness; content and cross-cultural validity	No eligible evidence	NE	Not graded
All	Structural validity of composite scores (reflective model only)	No eligible evidence	NE	Not graded

Note. R0 = original patient-level Mallet grade; V1–V7 = review-defined operationalisations detailed in Table 2. 3D = three-dimensional; AMS = Active Movement Scale; CV = construct validity; EMG = electromyography; ER = external rotation; GA = global abduction; GER = global external rotation; HN = hand to neck; HM = hand to mouth; HS = hand to spine; IC = internal consistency; IRR = inter-rater reliability; MIC = minimal important change; MRI = magnetic resonance imaging. + = sufficient; − = insufficient; ± = inconsistent; ? = indeterminate; NR = not rated; NA = not applicable; NE = no eligible evidence. Hypothesis-testing and non-rating criteria are described in Methods; certainty was graded only for rated results. Structural validity and IC apply only if a composite score is intended to reflect a single latent construct; under a formative model, they are not applicable.

4. Discussion

The findings of this systematic review indicate that "Mallet" does not refer to a single, homogeneous measurement instrument. Across the thirteen included studies, which included samples of 15 to 193 participants aged 1 month to 19 years (the youngest participants were assessed in a single reliability study [8]) and were published between 2003 and 2025, measurement performance depended on the operationalisation used, the task and scoring level assessed, the rater configuration, and the assessment setting. Reliability was sufficient for some tasks and inconsistent for others; construct-validity evidence was more consistent at the task level than at the composite level; and agreement with angle-defined motion analysis was insufficient wherever it was examined. No study addressed responsiveness, content validity, cross-cultural validity, or a minimal important change, and no multi-item composite form was evaluated for structural validity. Certainty of evidence was low or very low for almost every synthesis, reflecting both a small evidence base and methodological limitations in the contributing studies.

Seven review-defined Mallet operationalisations were identified. V1 comprised the five-task item-scored modified Mallet [8,28,29]; V2 substituted flexion for global external rotation within a non-standard five-task set [30]; V3 added hand-to-belly/internal rotation to produce a six-task form [9,14–16,31]; V4 added a separate supination task and an arm-at-rest classification [17]; V5 combined the six-task form with an arm-at-rest classification [27]; V6 reported the six-task form as a single overall grade without describing its derivation [32]; and V7 used only the global external-rotation item, assessed in shoulder adduction [33].

These operationalisations differ in task content, scoring unit, aggregation, and reporting structure. The original patient-level Mallet grade [10] was treated separately as reference form R0. Although its clinical applicability has been examined outside the included evidence base [13], no included study evaluated an eligible measurement property of that form. Evidence generated for one operationalisation or scoring level should therefore not be assumed to generalise to another. Because the V1–V7 boundaries were review-defined, we also examined whether merging the closely related extended forms (V3–V5) would alter the conclusions. It would not change the direction of any synthesis-level rating: V4 contributed only reliability evidence, for which no V3 or V5 evidence existed, and the V3 and V5 known-group findings were both sufficient.

Reliability within V1 varied by task rather than being uniformly strong or weak across the five-task set. In van der Sluijs et al. [28], hand-to-neck reliability was consistently sufficient, hand-to-mouth reliability was consistently insufficient, and global abduction, global external rotation, and hand-to-spine were inconsistent across observer pairs. In the single V4 study, Delioğlu et al. [17] found that reliability varied more across rater configurations than across assessment modes. Agreement between two trained raters was sufficient across face-to-face, live tele-, and recorded-video assessments ($\kappa = 0.79$–1.00); however, when four raters assessed the children remotely, only three of eight components were sufficient, while the remaining estimates ranged from $\kappa = 0.63$ to 0.65. These findings suggest that rater configuration and protocol standardisation may influence reliability, but they derive from a single low-certainty study and should not be taken to establish that rater effects are generally more important than assessment mode. One possible explanation for this task-specific variability is the nature of the construct assessed by the Mallet systems. Rather than evaluating isolated glenohumeral motion, Mallet tasks require coordinated contributions from the glenohumeral joint, scapulothoracic articulation, trunk, elbow, and forearm. Consequently, children may achieve similar functional positions through different compensatory movement strategies, introducing ambiguity into observer ratings, particularly for complex tasks such as hand-to-mouth. Thus, part of the observed variability may reflect genuine differences in movement strategies rather than observer inconsistency alone.

Favourable findings at the task level cannot be assumed to extend to total or composite scores. Bae et al. [8] reported high correlations for repeated aggregate scores, with mean r values of 0.92 within raters and 0.78 between raters. However, because these were Pearson correlations rather than intraclass correlation coefficients or weighted kappas, they did not meet the COSMIN criterion for sufficient reliability and were rated indeterminate; the reliability of aggregate Mallet scores therefore remains unresolved. The task-level breakdown needed to reconcile these findings with the more mixed single-task results was unavailable. No included study reported a standard error of measurement, smallest detectable change, limits of agreement, or minimal important change for any composite or total score, so no eligible evidence on measurement error was available. Cronbach's alpha, where reported, could not be meaningfully interpreted in the absence of evidence that the underlying items form a unidimensional construct, and no study examined the structural validity of any multi-item Mallet form used to calculate a composite or total score. Until these gaps are addressed, total and composite Mallet scores should be regarded as convenient summaries rather than as validated unidimensional measures. This reflects an absence of evidence rather than evidence against composite scores: no included study showed that the Mallet tasks are multidimensional or that summation is inappropriate. Composite scores may legitimately serve as global summaries of shoulder function, but which properties they require depends on their intended measurement model. If the total is intended to reflect a single latent shoulder-function construct, its dimensionality should be tested with factor-analytic or Rasch methods before internal consistency is interpreted; if it is intended

as a formative index of distinct functional tasks, structural validity and internal consistency are not the relevant criteria, and the adequacy of the task set should instead be established through content-validity studies. Until either route has been followed, the dimensionality and measurement properties of composite Mallet scores remain unestablished.

Construct-validity evidence was more consistent when the comparator assessed the same movement than when it assessed a different level of function. All six Mallet–Active Movement Scale task/score comparisons met expectations [9,12], as did all five shoulder-strength correlations [30] and six of eight task-level correlations with three-dimensional kinematics [14]. At the composite level, however, only two of three comparisons with other clinical scales met expectations, with the relationship to the Toronto Test Score falling short [9]. Evidence linking Mallet scores to muscle-level comparators was mixed, with infraspinatus volume but not electromyographic interference grade meeting expectations [33]. This pattern suggests that Mallet tasks capture specific observable functional performances and correlate most strongly with closely related, same-movement measures. Associations weaken when comparators represent different levels of the International Classification of Functioning, Disability and Health or composite indices combining distinct movement patterns.

Two studies compared clinical Mallet grading with three-dimensional motion capture, and both found insufficient agreement, with clinical scores systematically overestimating movement, particularly for external rotation. Russo et al. [16] found exact agreement for external rotation in only 6 of 26 participants, with every disagreement favouring the clinical grade. Lovette et al. [15] reported near-zero agreement for external rotation ($\kappa = 0.065$), moderate agreement for global abduction ($\kappa = 0.50$), and low classification accuracy for hand-to-mouth despite reasonable sensitivity, with mean clinical overgrading of 1.03 grades for external rotation and 0.17 grades for abduction. A plausible explanation is that trunk and scapular compensations, which are difficult to fully discount through visual observation, may contribute to the assigned clinical grade. This has a direct interpretive consequence: a Mallet grade should not be treated as a proxy for angle-defined glenohumeral motion, and clinicians and researchers who need an angle-defined measure of isolated glenohumeral motion should not substitute Mallet grading for instrumented motion analysis. This also means that improvements in Mallet grades following surgery or rehabilitation should not automatically be interpreted as restoration of normal shoulder biomechanics. They may instead reflect improved motor control, more efficient compensatory strategies, or enhanced task performance despite persistent biomechanical impairment. Equally, poor agreement with motion capture should not be read as evidence that the Mallet tasks are invalid. Instrumented motion analysis quantifies segmental joint angles, whereas Mallet grading records whether a functional end position is achieved, including any compensatory contribution; the two methods therefore address related but distinct constructs. The observed disagreement indicates that Mallet grades cannot substitute for angle-based measurement, not that they fail to capture functional task performance.

Nevertheless, these limitations should not be interpreted as indicating that the Mallet classification systems lack clinical value. Rather, they highlight the need for appropriate interpretation of the scores according to the specific version and task assessed. Despite the limitations identified in this review, the Mallet systems remain widely used clinician-reported observational measures of shoulder function in children with BPBI and are included among the core outcome measures recommended by the iPLUTO consensus [7]. Our findings qualify this recommendation: the available evidence is property-, task-, and operationalisation-specific, is mostly of low or very low certainty, and does not demonstrate the overall superiority of any Mallet form. Instrumented motion analysis provides complementary information when precise quantification of isolated glenohumeral motion is

required; however, it is not a practical substitute for routine clinical assessment. Until more robust observational instruments with stronger psychometric evidence become available, careful version-specific use of the Mallet systems appears justified in both clinical practice and longitudinal follow-up.

Taken together, these findings support using the Mallet classification systems as task-based, clinician-reported observational measures that complement, rather than replace, other outcome measures and instrumented assessment in brachial plexus birth injury. Task-level scores may be useful for describing specific functional movements, particularly when raters are trained and scoring is standardised; however, their ability to detect or quantify longitudinal change remains unknown. Composite and total scores currently rest on a substantially thinner evidence base. As a provisional suggestion, because the supporting evidence is of low or very low certainty and derives from few small studies, users of the V1 five-task item-scored modified Mallet may report individual task scores obtained by trained raters with a standardised protocol, together with the task set, rater configuration, and scoring protocol; this suggestion should be revisited as higher-quality evidence becomes available. However, these scores should not be assumed to detect or quantify true longitudinal change because measurement error, responsiveness, and minimal important change remain unknown. Composite or total scores should be interpreted cautiously and, where possible, accompanied by the underlying task-level breakdown rather than reported alone. Future research should prioritise: (1) standardising the reporting of Mallet versions and task sets; (2) establishing content validity with input from clinicians, children, and caregivers; (3) specifying the intended measurement model of composite scores and, where a reflective model is intended, evaluating structural validity using factor-analytic or Rasch methods; and (4) conducting reliability and measurement-error studies using appropriate agreement statistics, alongside studies of responsiveness and minimal important change.

This review's principal strength is that it is, to our knowledge, the first to apply COSMIN methodology to the Mallet classification systems while explicitly distinguishing between versions, tasks, and scoring levels rather than treating "Mallet" as a single instrument; this stratified approach made it possible to identify where the evidence is reasonably consistent and where it is not. The review is limited by the small number of eligible studies (thirteen), the doubtful or inadequate risk-of-bias ratings for most construct-validity analyses, and the complete absence of evidence for measurement error, responsiveness, content validity, structural validity, and cross-cultural validity. Several contributing samples may overlap, particularly among the Shriners motion-capture reports and the Hacettepe reports, which could not be fully resolved from the published information. Several methodological decisions also warrant caution. COSMIN standards developed primarily for patient-reported outcome measures were applied to a clinician-rated observational instrument; although the quantitative criteria were retained, their suitability for ordinal clinician-rated tasks has not been formally established, and the thresholds for agreement with an angle-defined reference were review-specific. The construct-validity hypotheses were not included in the registered protocol and were formulated after data extraction, when the extracted results were known; although hypothesis classes were assigned on conceptual grounds and COSMIN thresholds were used, knowledge of the results may have influenced hypothesis specification. Operationalisation codes were assigned jointly by two reviewers rather than independently, and extraction, risk-of-bias, sufficiency, and certainty judgements were made by one reviewer and verified by a second, rather than being performed independently in duplicate. Five of the 13 included studies were identified only through supplementary searching, suggesting that measurement-property search filters may miss eligible analyses embedded in clinical, surgical, or biomechanical reports; an additional Embase search covering the same period identified 52 records, 7 of which were unique after deduplication, and none

were eligible. In addition, the measurement model of composite Mallet scores could not be determined from the available evidence; conclusions on their structural validity and internal consistency are therefore conditional on the intended use of these scores. The exclusion of non-English full-text publications may also have introduced language bias. Meta-analysis was not feasible given the heterogeneity in versions, scoring levels, rater designs, and comparators, so findings were synthesised narratively, and most graded syntheses carried low or very low certainty. These limitations mean that the conclusions reflect the current, fragmented state of the evidence rather than definitive judgements about the Mallet systems' measurement properties.

5. Conclusions

The Mallet classification systems provide limited, operationalisation-specific evidence on shoulder function in children with brachial plexus birth injury, most of which is of low or very low certainty. Task-level findings were more consistent than composite-score findings, but no eligible evidence was identified for measurement error, responsiveness, or minimal important change; Mallet scores therefore cannot currently be regarded as validated measures of longitudinal change. The dimensionality and measurement properties of composite scores remain unestablished, and equivalence across operationalisations has not been demonstrated. Any preference for V1 task-level reporting should be regarded as provisional. Until further evidence is available, Mallet results are best interpreted as complementary clinician-reported outcomes, reported explicitly by operationalisation, task set, and scoring level, and not as a substitute for instrumented motion analysis.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/children13111604/s1; Table S1: Study-to-operationalisation mapping; Table S2: Full search strategies; Table S3: Screening results with exclusion reasons; Table S4: Construct-validity hypotheses and rating rules; Table S5: Result-level evidence table; Table S6: Evidence synthesis map.

Author Contributions: Conceptualization, A.Ö.A., B.C., and Z.H.; methodology, A.Ö.A. and B.C.; validation, Z.H.; investigation, A.Ö.A. and B.C.; data curation, A.Ö.A. and B.C.; writing—original draft preparation, A.Ö.A. and B.C.; writing—review and editing, A.Ö.A., B.C., and Z.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new primary data were generated in this study. The search strategies, study-to-operationalisation mapping, screening decisions, construct-validity hypotheses, result-level extraction and rating data, and the evidence synthesis map supporting the reported findings are provided in Tables S1–S6.

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT (OpenAI; GPT-5 models, accessed July 2026) for language refinement and to assist with the development and refinement of R code used to generate Figures. The authors reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BPBI	Brachial plexus birth injury
ClinROMs	Clinician-reported observational measures
COSMIN	Consensus-based Standards for the selection of health Measurement Instruments
GRADE	Grading of Recommendations Assessment, Development and Evaluation
iPLUTO	International Plexus Outcome Study Group
OMI	Outcome measurement instrument
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-COSMIN for OMIs	PRISMA–COSMIN guideline for systematic reviews of outcome measurement instruments
PROM	Patient-reported outcome measure
PROSPERO	International Prospective Register of Systematic Reviews
R0	Original patient-level Mallet grade used as the historical reference form
V1–V7	Review-defined Mallet operationalisations based on task content and scoring or reporting structure

References

1. Dorich, J.M.; Whiting, J.; Plano Clark, V.L.; Ittenbach, R.F.; Cornwall, R. Impact of brachial plexus birth injury on health-related quality of life in adulthood: A mixed methods survey study. *Disabil. Rehabil.* **2024**, *46*, 2042–2055. [CrossRef] [PubMed]
2. Waters, P.M. Comparison of the natural history, the outcome of microsurgical repair, and the outcome of operative reconstruction in brachial plexus birth palsy. *J. Bone Jt. Surgery. Am. Vol.* **1999**, *81*, 649–659. [CrossRef] [PubMed]
3. Duff, S.V.; DeMatteo, C. Clinical assessment of the infant and child following perinatal brachial plexus injury. *J. Hand Ther.* **2015**, *28*, 126–134. [CrossRef] [PubMed]
4. Khabyeh-Hasbani, N.; O'Brien, D.M.; Meisel, E.M.; Koehler, S.M. Current concepts in brachial plexus birth injuries: A comprehensive narrative review. *Plast. Reconstr. Surg.–Glob. Open* **2024**, *12*, e6083. [CrossRef] [PubMed]
5. Manske, M.C.; Abarca, N.E.; Letzelter, J.P.; James, M.A. Patient-Reported Outcomes Measurement Information System (PROMIS) scores for children with brachial plexus birth injury. *J. Pediatr. Orthop.* **2021**, *41*, 171–176. [CrossRef] [PubMed]
6. Chang, K.W.; Justice, D.; Chung, K.C.; Yang, L.J.S. A systematic review of evaluation methods for neonatal brachial plexus palsy: A review. *J. Neurosurg. Pediatr.* **2013**, *12*, 395–405. [CrossRef] [PubMed]
7. Pondaag, W.; Malessy, M.J.A. Outcome assessment for brachial plexus birth injury: Results from the iPluto world-wide consensus survey. *J. Orthop. Res.* **2018**, *36*, 2533–2541. [CrossRef] [PubMed]
8. Bae, D.S.; Waters, P.M.; Zurakowski, D. Reliability of three classification systems measuring active motion in brachial plexus birth palsy. *J. Bone Jt. Surgery. Am. Vol.* **2003**, *85*, 1733–1738. [CrossRef] [PubMed]
9. Greenhill, D.A.; Lukavsky, R.; Tomlinson-Hansen, S.; Kozin, S.H.; Zlotolow, D.A. Relationships between 3 classification systems in brachial plexus birth palsy. *J. Pediatr. Orthop.* **2017**, *37*, 374–380. [CrossRef] [PubMed]
10. Mallet, J. Paralysie obstétricale du plexus brachial. Traitement des séquelles: Primauté du traitement de l'épaule—Méthode d'expression des résultats. *Rev. Chir. Orthopédique Réparatrice l'Appareil Mot.* **1972**, *58*, 166–168.
11. Sarıipek, M.; Seven, O.F.; Yalçın, A.İ.; Topuz, S.; Fırat, T. The relative contribution of trunk movements to Modified Mallet Scores in children with brachial plexus birth injury. *J. Shoulder Elb. Surg.* **2026**, *35*. [CrossRef] [PubMed]
12. Curtis, C.; Stephens, D.; Clarke, H.M.; Andrews, D. The Active Movement Scale: An evaluative tool for infants with obstetrical brachial plexus palsy. *J. Hand Surg.* **2002**, *27*, 470–478. [CrossRef] [PubMed]
13. Al-Qattan, M.M.; El-Sayed, A.A. Obstetric brachial plexus palsy: The mallet grading system for shoulder function–revisited. *BioMed Res. Int.* **2014**, *2014*, 398121. [CrossRef] [PubMed]
14. Russo, S.A.; Kozin, S.H.; Zlotolow, D.A.; Nicholson, K.F.; Richards, J.G. Motion necessary to achieve Mallet internal rotation positions in children with brachial plexus birth palsy. *J. Pediatr. Orthop.* **2019**, *39*, 14–21. [CrossRef] [PubMed]
15. Lovette, M.; Chafetz, R.S.; Russo, S.A.; Kozin, S.H.; Zlotolow, D.A. Shoulder motion overestimated by Mallet scores. *J. Pediatr. Orthop.* **2024**, *44*, e951–e956. [CrossRef] [PubMed]
16. Russo, S.A.; Chafetz, R.S.; Rodriguez, L.M.; Roposh, C.M.; Zlotolow, D.A.; Kozin, S.H.; Gaughan, J.P.; Richards, J.G. Comparison of shoulder motion measurements by visual estimate, goniometer, and motion capture. *J. Pediatr. Orthop.* **2022**, *42*, 443–450. [CrossRef] [PubMed]

17. Delioğlu, K.; Unes, S.; Tuncdemir, M.; Ozal, C.; Bıyık, K.S.; Uzumcugil, A. Interrater reliability of face-to-face, tele- and video-based assessments with the modified Mallet classification in brachial plexus birth injuries. *J. Hand Surg. Eur. Vol.* **2024**, *49*, 576–582. [CrossRef] [PubMed]
18. Mokkink, L.B.; Elsman, E.B.M.; Terwee, C.B. COSMIN guideline for systematic reviews of patient-reported outcome measures version 2.0. *Qual. Life Res.* **2024**, *33*, 2929–2939. [CrossRef] [PubMed]
19. Prinsen, C.A.C.; Mokkink, L.B.; Bouter, L.M.; Alonso, J.; Patrick, D.L.; de Vet, H.C.W.; Terwee, C.B. COSMIN guideline for systematic reviews of patient-reported outcome measures. *Qual. Life Res.* **2018**, *27*, 1147–1157. [CrossRef] [PubMed]
20. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [CrossRef] [PubMed]
21. Elsman, E.B.M.; Mokkink, L.B.; Terwee, C.B.; Beaton, D.; Gagnier, J.J.; Tricco, A.C.; Baba, A.; Butcher, N.J.; Smith, M.; Hofstetter, C.; et al. Guideline for reporting systematic reviews of outcome measurement instruments (OMIs): PRISMA-COSMIN for OMIs 2024. *J. Clin. Epidemiol.* **2024**, *173*, 111422. [CrossRef] [PubMed]
22. Mokkink, L.B.; Terwee, C.B.; Patrick, D.L.; Alonso, J.; Stratford, P.W.; Knol, D.L.; Bouter, L.M.; de Vet, H.C.W. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J. Clin. Epidemiol.* **2010**, *63*, 737–745. [CrossRef] [PubMed]
23. Mokkink, L.B.; Boers, M.; van der Vleuten, C.P.M.; Bouter, L.M.; Alonso, J.; Patrick, D.L.; de Vet, H.C.W.; Terwee, C.B. COSMIN Risk of Bias tool to assess the quality of studies on reliability or measurement error of outcome measurement instruments: A Delphi study. *BMC Med. Res. Methodol.* **2020**, *20*, 293. [CrossRef] [PubMed]
24. Mokkink, L.B.; de Vet, H.C.W.; Prinsen, C.A.C.; Patrick, D.L.; Alonso, J.; Bouter, L.M.; Terwee, C.B. COSMIN Risk of Bias checklist for systematic reviews of patient-reported outcome measures. *Qual. Life Res.* **2018**, *27*, 1171–1179. [CrossRef] [PubMed]
25. Terwee, C.B.; Jansma, E.P.; Riphagen, I.I.; de Vet, H.C.W. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. *Qual. Life Res.* **2009**, *18*, 1115–1123. [CrossRef] [PubMed]
26. Ouzzani, M.; Hammady, H.; Fedorowicz, Z.; Elmagarmid, A. Rayyan—A web and mobile app for systematic reviews. *Syst. Rev.* **2016**, *5*, 210. [CrossRef] [PubMed]
27. Delioğlu, K.; Öztürk, E.; Üzümcügil, A.; Günel, M.K. Investigation of upper extremity resting position in children with brachial plexus birth injury. *J. Hacet. Univ. Phys. Ther. Rehabil. Fac.* **2025**, *3*, 39–47.
28. van der Sluijs, J.A.; van Doorn-Loogman, M.H.; Ritt, M.J.P.F.; Wuisman, P.I.J.M. Interobserver reliability of the Mallet score. *J. Pediatr. Orthop. Part B* **2006**, *15*, 324–327. [CrossRef] [PubMed]
29. Bae, D.S.; Waters, P.M.; Zurakowski, D. Correlation of Pediatric Outcomes Data Collection Instrument with measures of active movement in children with brachial plexus birth palsy. *J. Pediatr. Orthop.* **2008**, *28*, 584–592. [CrossRef] [PubMed]
30. Brochard, S.; Alter, K.; Damiano, D.L. Shoulder strength profiles in children with and without brachial plexus palsy. *Muscle Nerve* **2014**, *50*, 60–66. [CrossRef] [PubMed]
31. Russo, S.A.; Topley, M.T.; Richardson, R.T.; Richards, J.G.; Chafetz, R.S.; Rapp van Roden, E.A.; Zlotolow, D.A.; Mulcahey, M.J.; Kozin, S.H. Assessment of the relationship between Brachial Plexus Profile activity short form scores and modified Mallet scores. *J. Hand Ther.* **2022**, *35*, 51–57. [CrossRef] [PubMed]
32. Ismaeel, M.M.I.; El-Saeed, T.M.; Hafez, E.A. Spinal deviations in relation to arm functions in obstetrical brachial plexus injuries. *Life Sci. J.* **2014**, *11*, 1077–1081.
33. Talbert, R.J.; Michaud, L.J.; Mehlman, C.T.; Kinnett, D.G.; Laor, T.; Foad, S.L.; Schnell, B.; Salisbury, S. EMG and MRI are independently related to shoulder external rotation function in neonatal brachial plexus palsy. *J. Pediatr. Orthop.* **2011**, *31*, 194–204. [CrossRef] [PubMed]
34. Michelow, B.J.; Clarke, H.M.; Curtis, C.G.; Zuker, R.M.; Seifu, Y.; Andrews, D.F. The natural history of obstetrical brachial plexus palsy. *Plast. Reconstr. Surg.* **1994**, *93*, 675–680. [CrossRef]
35. Narakas, A.O. Obstetrical brachial plexus injuries. In *The Paralysed Hand*; Lamb, D.W., Ed.; Churchill Livingstone: Edinburgh, Scotland, 1987; pp. 116–135.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.