ASHA American Speech-Language-Hearing Association

AJSLP

Clinical Focus

From Foundations to Frontiers: Fit-for-Purpose Discourse Science

Brielle C. Stark[a]

[a] Department of Speech, Language and Hearing Sciences, Indiana University Bloomington

ARTICLE INFO

Article History:
Received February 3, 2026
Revision received April 9, 2026
Accepted April 28, 2026

Editor-in-Chief: Rita R. Patel
Editor: Mary R. T. Kennedy

https://doi.org/10.1044/2026_AJSLP-26-00055

ABSTRACT

Purpose: This clinical focus article responds to discussions at the 2026 International Cognitive-Communication Disorders Conference (ICCDC), where participants highlighted a productive tension that has shaped discourse research for decades. Canonical discourse elicitation tasks (e.g., Cinderella retell) remain valuable because they support historical continuity, cross-study comparison, and shared data sets, yet they do not always capture the discourse constructs most relevant for all populations or clinical purposes. Discourse analysis has long been central to the assessment and treatment of neurogenic and cognitive-communication disorders, and canonical tasks have shaped decades of research by yielding replicable, analyzable, and linguistically rich language samples.
Method: This article is a narrative review that draws on the author's personal experience as a translational researcher and deep knowledge of the field.
Results and Conclusions: This article examines what canonical tasks revealed about language breakdown, why they succeeded within their original scope, where they may be inappropriate, and why and when they should still be retained. Building on discussion from the ICCDC, this piece argues that the field does not simply need "new tasks" but rather a principled framework for deciding when canonical tasks remain useful; when they should be adapted; and when new tasks, methods, and metrics are warranted. A fit-for-purpose discourse science is proposed, in which discourse tasks and outcome measures are selected because they are appropriate; robust; and aligned with the purpose, population, and construct of interest, while also integrating psychometric rigor, participatory design, and thoughtful use of technologies such as natural language processing and perceptual ratings. This clinical focus article also offers practical tables summarizing elicitation methods, suitable measures, key features of strong discourse assessment, clinically feasible implementation principles, and example scenarios for choosing canonical and noncanonical tasks in research and practice.

This clinical focus article is from the perspective of a researcher whose work centers on discourse sampling and analysis in neurogenic and cognitive-communication disorders, with research spanning aphasia, traumatic brain injury (TBI), right-hemisphere disorder (RHD), aging, and progressive neurodegenerative diseases. Over the past decade, my work has examined how discourse elicitation tasks shape language, impact psychometric properties, and influence clinical interpretability (B. C. Stark, 2019, 2023; B. C. Stark et al., 2021, 2022, 2023). These insights drive this piece's central argument: Discourse assessment remains essential, but its future requires greater clarity about what canonical tasks taught us, where their limitations lie, and how to achieve fit-for-purpose discourse assessment.

Method and Scope

This clinical focus article is a narrative review that draws on the author's personal experience as a translational researcher and deep knowledge of the field. It argues for a shift from tradition-based task selection toward approaches

Correspondence to Brielle C. Stark: bcstark@iu.edu. **Publisher Note:** This article is part of the Special Issue: Select Papers From the Fifth International Cognitive-Communication Disorders Conference. ***Disclosure:*** *The author has declared that no competing financial or nonfinancial interests existed at the time of publication.*

that are intentional, theory driven, psychometrically robust, and empirically validated.

A Brief Introduction to Clinical Discourse Assessment and Analysis

Discourse sampling captures language in use—extended, self-generated, and contextually grounded—making it particularly valuable in neurogenic and cognitive-communication disorders. Unlike isolated word or sentence tasks, discourse requires the integration of multiple levels of linguistic and cognitive processing, including lexical retrieval, syntax, coherence, pragmatic appropriateness, and executive control. Monologic discourse tasks such as picture description or narrative retell allow for greater experimental control than conversation, reducing variability introduced by partner interaction, turn-taking, or topic negotiation. This helps explain why canonical discourse tasks (to be discussed here) have been monologic. At the same time, this experimental control should not diminish the importance of co-constructed and conversational discourse elicitation, which remains essential for understanding the pragmatic, interactive, and participation-level aspects of communication that monologic tasks may miss. For researchers, discourse tasks provide a window into the neural and cognitive underpinnings of language. For clinicians, discourse tasks offer a more ecologically valid view of how language breakdowns affect daily communication.

Table 1 provides an overview of common discourse elicitation types: what they elicit, common metrics, strengths, limitations/cautions, and other considerations (such as for whom [populations, impairments, goals] they may be appropriate). Table 2 summarizes key features of a strong discourse assessment from a researcher's point of view, offering a practical framework for selecting tasks and metrics and planning a robust study. While not every feature in Table 2 is likely to be feasible in routine clinical scenarios, Table 3 summarizes actionable clinical directions for discourse assessment. Table 3 also complements recent tutorials that provide practical guidance for making discourse assessment and treatment more usable in cognitive-communication and language therapy (Dutta et al., 2025; Leaman & Archer, 2023).

A Rich History

Two canonical (monologic) tasks—Cookie Theft picture description and Cinderella narrative retell—have become foundational in clinical and research settings. The Cookie Theft task, released as part of the Boston Diagnostic Aphasia Examination (Goodglass & Kaplan, 1972), uses a detailed scene to elicit connected speech. It features multiple agents, foreground/background elements, simultaneous actions, and potential emotional stakes ("the boy falling!" "the water overflowing!"), all of which support the elicitation of sentence structures, connectors, object/action naming, and pronouns. The visual scaffold reduces cognitive load, supporting speakers with memory or planning challenges. The constrained content and shared knowledge between the clinician and the speaker minimize nonlinguistic variability and improve the reliability of linguistic measurement. Clinicians can more easily identify errors and analyze specific features such as speech rate, fluency, referential specificity, and lexical retrieval. However, the task has also been critiqued for being outdated and culturally narrow, including depictions of traditional gender roles, prompting calls for updated versions. In response, a modern Cookie Theft picture was recently created; the revised color stimulus modernizes the scene by depicting a man washing dishes while children reach for cookies and by adding a woman mowing the lawn while talking on a cell phone, a cat chasing birds, and a dog in the kitchen (Berube et al., 2019), though only recently have psychometric and normative data been published (Dalton et al., 2024; Hux & Frodsham, 2023), and there are limited data on its use in other populations, especially in cognitive-communication disorders (Berube et al., 2022).

The Cinderella narrative retell has similarly shaped discourse analysis, at least as early as 1987 in aphasia (Hadar et al., 1987) and 1991 in a person with right-hemisphere temporal lobe hematoma (Hadar et al., 1991). Participants typically view a wordless picture book version of Cinderella (e.g., Grimes, 2005) and are asked to retell the story, often without visual cues during the narrative portion. This format supports extended language (often many more tokens than produced during picture descriptions [B. C. Stark, 2019; B. C. Stark & Fukuyama, 2021], such as the Cookie Theft task) and provides a scaffold for examining other critical aspects of discourse, including coherence, sequencing, and narrative macrostructure (e.g., story grammar, content elements; Greenslade et al., 2020; J. A. Stark, 2010). Cinderella encourages propositionally dense language and has been used to differentiate aphasia subtypes and to assess planning, sequencing, pragmatic, and coherence impairments in other neurogenic and cognitive-communication conditions (Bartels-Tobin & Hinckley, 2005; Greenslade et al., 2024). Because Cinderella is a highly familiar story for many U.S. participants, that familiarity can reduce demands on novel encoding and support the analysis of narrative organization and content expression. However, familiarity can also blunt sensitivity: Stockbridge and Newman (2019) found no group differences for Cinderella retell in adults with concussion, whereas differences emerged for a previously unknown narrative, suggesting that known and novel retells probe partly different cognitive–linguistic demands. Yet, similar to the Cookie Theft task, the Cinderella task has clear limitations. The story's infantilizing, gendered themes

Table 1. Common discourse elicitation types: what they elicit, common metrics, strengths, limitations/cautions, and other considerations.

Elicitation type	Typical example	What it tends to elicit	Common metrics well suited to this task	Strengths	Limitations/cautions	Often useful for
Picture description	• Cookie Theft • Cat Rescue (TalkBank) • Picnic Scene (Western Aphasia Battery) • Other complex scenes	• Constrained descriptive language • Object/action naming • Sentence-level production • Referential language (e.g., connecting actors to actions, to other actors)	• Word/token count • Informativeness • Main concepts • Core lexicon • Lexical diversity • Efficiency of word/content/main concept production • Referential clarity	• High structure • Low memory burden • Easy to administer • Good cross-speaker comparability • Often strong replicability	• May under-sample extended discourse • Tends to elicit simple syntax • Limited ecological validity • Cultural or historical bias in stimuli • May miss pragmatic or higher level discourse problems	• Disorders where language is more heavily impacted (e.g., aphasia) • Quick characterization of microlinguistic abilities • Studies needing controlled content
Story/narrative retell	• Cinderella • Wordless picture book retell (e.g., "Good Dog Carl") • Wordless video retell	• Longer connected speech • Temporal/causal sequencing • Highly referential language	• Story grammar • Global/local coherence • Core lexicon • Main concepts • Informativeness • Proposition density • Syntactic complexity • Lexical measures • Topic maintenance • Gist	• Elicits richer discourse than many constrained tasks • Supports analysis of macro- and microlinguistic features	• Prior familiarity can inflate performance • Unfamiliar or culturally distant stories may disadvantage some speakers • Memory load can affect performance	• Aphasia, TBI, RHD, dementia • Studies of organization, narrative structure, and discourse richness
Procedural discourse	"Tell me how to make a sandwich" or "how to mail a letter"	• Ordered, directive language • Concise stepwise sequencing • Associates with relatively simple syntax	• Informativeness • Sequencing accuracy • Efficiency • Core lexicon • Main concepts • Sentence structure • Representational gesture (tends to elicit iconic gestures)	• Familiar everyday content/functional • Clinically relevant • Useful for instructional communication	• Can yield short/simple output • Lower linguistic richness than narratives • Experience with procedure varies across speakers	• Evaluating nonverbal and functional communication • Executive sequencing demands

(*table continues*)

Table 1. *(Continued).*

Elicitation type	Typical example	What it tends to elicit	Common metrics well suited to this task	Strengths	Limitations/cautions	Often useful for
Personal narrative	"Tell me about an important event" (TalkBank)	• Self-generated, autobiographical, (sometimes) emotionally meaningful discourse	• Local/global coherence • Informativeness • Lexical diversity • Referential clarity • Pragmatic appropriateness • Listener ratings • Narrative structure	• High ecological validity • Engaging and often motivating • Can reveal real-world communication success/failure	• Harder to standardize • Topic familiarity and emotional salience vary. • Lower replicability across speakers/time • Some pragmatic ability to create common ground with listener is required if a shared context is not there (listener does not know the story details already)	• Functional assessment • Person-centered clinical work • Discourse participation goals
Conversation/interview/co-constructed communication	• Semistructured interview • Spontaneous conversation • Semistructured conversation (a priori topic selection) • Cooperative gameplay or narrative	• Turn-taking • Topic management • Local/global coherence • Conversational and turn repair • Pragmatic behavior (e.g., adapting to listener, question asking)	• Conversation analysis • Pragmatic ratings • Communicative success ratings • Local/global coherence • Turn-taking • Repair frequency • Listener judgment	• Reflective of real-world communication • Captures interactive competence and social appropriateness	• Lowest experimental control • Substantial partner effects • More difficult cross-study comparability • Scoring burden can be high	• Pragmatics and social communication • Participation-level outcomes • May be most interesting for the cognitive-communication population

Note. TBI = traumatic brain injury; RHD = right-hemisphere disorder.

Table 2. Key features of a strong discourse science: a research framework.

Feature to consider	What it means in discourse assessment	Questions to ask when choosing a task and a metric	Red flags if not addressed	Practical takeaway
Purpose fit	The task–metric pair matches the intended goal.	Is the goal (a) impairment description, (b) differential profiling, (c) treatment planning, or (d) change detection?	Familiar tasks used by default without a stated rationale.	Start with the question being asked, then choose the task and metric.
Population fit	Demands align with the communication and cognitive profile of the target group.	Does the task fit aphasia, TBI, RHD, dementia, and/or healthy aging? Could the task miss the disorder's hallmark difficulties?	A task works in one population but is assumed to generalize to another.	Validate tasks and metrics in the population in which they will be used.
Construct validity	The task and the metric capture the intended discourse construct.	Does the metric reflect lexical retrieval, coherence, pragmatics, macrostructure, or functional success as intended?	Vague claims that a measure reflects "discourse ability" broadly.	Name the construct explicitly and justify the chosen metric by evaluating whether evidence supports using that task to evaluate your metric of interest.
Reliability	Scores are stable enough across raters, occasions, or samples.	Is there evidence for test–retest reliability, interrater/intrarater reliability, or stability across similar samples?	Measures shift substantially because of task, scorer, or sample length.	Do not assume that discourse measures or raters are stable; demonstrate it using within-subject or test–retest designs whenever possible, alongside adequate rater training in the outcome measure (e.g., metric).
Sensitivity to change	The measure can detect meaningful improvement or decline.	Would this task–metric pair plausibly show treatment effects or longitudinal change?	A task is used to derive a metric as an outcome measure without evidence that it can track change.	Outcome measures need responsiveness, not just descriptive value. Look for literature showing longitudinal evidence of metric change; if unavailable, create your own via strong within-subject design.
Replicability/ standardization	Administration, scoring, and analysis can be reproduced across users and sites.	Are prompts, instructions, transcription rules, and scoring procedures clearly documented and made openly available?	"Homegrown" procedures are used but not described in enough detail to repeat or are behind paywalls/closed websites.	Standardized instructions and transparent scoring improve comparability and are easily found on lab websites, as part of open repositories, and so forth.
Clinical utility/feasibility	The method is workable in real practice.	How much time, training, transcription, and scoring burden does it require?	Valuable in theory but too burdensome for routine clinical use.	Favor methods that balance rigor with real-world feasibility, such as perceptual methods that have been shown to be reliable clinically or automated methods implemented in the clinic.
Ecological relevance	The discourse sample reflects communication that matters outside the clinic or lab.	Does the task resemble meaningful everyday communication demands?	A highly controlled task is treated as fully representative of daily communication.	Use more than one task type when functional relevance is important. It is well acknowledged that no task perfectly resembles everyday communication; use a few tasks to get a better "snapshot" of communicative competence.

(*table continues*)

Table 2. *(Continued).*

Feature to consider	What it means in discourse assessment	Questions to ask when choosing a task and a metric	Red flags if not addressed	Practical takeaway
Interpretability in the context of client and treatment goals	Results can be explained in a way clinicians and researchers can use.	Does a score have a clear meaning in the context of client and treatment goals? Is "better" performance obvious and clinically meaningful?	Numbers are reported without explaining what they indicate about communication.	Prefer metrics that support actionable interpretation. For example, is "more" words always better? Or is being "verbose" not desirable? Choose metrics with clear interpretability that fits client and treatment goals.
Task–metric compatibility	The chosen metric makes sense for the discourse elicited (see Table 1).	Does the task reliably produce enough of the target behavior for the metric to be meaningful?	Macrostructure scored on a task with minimal narrative content; coherence scored on very short samples.	Strong metrics can underperform when paired with the wrong task, limiting or causing incorrect interpretation of the outcome.
Cultural and contextual appropriateness	Content, assumptions, and prompts fit the speaker's background and treatment goals.	Are the stimuli familiar, respectful, and relevant to the individual or community and goals?	Older canonical tasks are used despite cultural mismatch or alienating content.	Select or adapt prompts with inclusivity and relevance in mind.
Compatibility with automation and human judgment	The method can benefit from technology without losing clinical insight.	Can parts of the analysis be automated? Which features still require trained human judgment?	Automated scores are treated as sufficient without human "checking."	Hybrid approaches are often strongest: automation for efficiency, humans for interpretation.

Note. TBI = traumatic brain injury; RHD = right-hemisphere disorder.

can be alienating or culturally irrelevant, particularly in non-Western contexts. Familiarity with the plot may also mask deficits in individuals with mild impairments, whereas those unfamiliar with the story may struggle due to mismatched background assumptions.

Canonical discourse tasks, such as Cookie Theft and Cinderella, were adopted because they offered structured but linguistically rich prompts, reduced nonlinguistic variability, supported consistent administration, and enabled shared data sets and cross-study comparison. These strengths made them especially valuable for impairment characterization and helped establish common reference points for the field, including major corpora such as AphasiaBank, TBIBank, RHDBank, and DementiaBank (MacWhinney, 2007). Their success, however, was tied to that original purpose: Highly structured tasks are well suited to profiling language impairment but are not automatically optimal for individualized treatment planning, ecologically grounded assessment, or sensitive measurement of change. Their limitations become most evident when they are expected to capture pragmatic deficits in cognitive-communication disorders, reflect culturally diverse lived experiences, support treatment-sensitive outcomes, or represent everyday communicative success, even as some updated stimuli have begun to address aspects of these concerns. Thus, the argument is not that canonical tasks are now inherently invalid or obsolete. Rather, their enduring value lies in their structure, familiarity, analyzability, and comparability across participants, sites, and decades of research. The problem arises when they are treated as universally appropriate, or when the metrics derived from them are assumed to be valid for every population, construct, and clinical purpose.

The Influence of Task on Language and Its Organization During Discourse

It is important to distinguish between *discourse elicitation tasks* and the *metrics* used to analyze the discourse they generate (see Table 1). Elicitation tasks—such as picture descriptions, story retells, or conversational interviews—are tools designed to prompt language samples. Metrics, by contrast, are the quantitative or qualitative measures applied to those samples to assess linguistic, cognitive, or communicative features. Although these two components work in tandem, each poses distinct methodological challenges and requires separate validation. Conflating them risks obscuring important differences in how discourse is collected versus how it is analyzed and interpreted.

The importance of task design cannot be overstated. Task choice shapes not only the language that is elicited but also the psychometric properties of the measures derived from that language. Different discourse genres

Table 3. Key features of a strong discourse science: a clinical framework.

Clinical priority	What is feasible in everyday practice	What to avoid	Link to Tables 1 and 2
Start with the communication purpose.	Choose tasks based on the client's real-world needs and the clinical question, such as describing, telling, explaining, or conversing.	Using the same familiar task for every client regardless of goals.	Reflects **purpose fit** in Table 2 and the different **elicitation demands** shown in Table 1.
Match the task to the person and population, to the best of your ability, with what you have.	Select one or two tasks the client can engage in meaningfully, considering diagnosis, cognitive demands, cultural fit, background knowledge, and goals.	Assuming that a task that works for one client—or in one diagnosis or in research—will work equally well for every client.	Reflects **population fit** and **cultural/contextual appropriateness** in Table 2 and uses **genre differences** in Table 1.
Use task–metric pairs that make sense together.	Pick measures that fit the discourse sample, such as informativeness for picture description, main concepts or sequencing for narrative, or communicative success for conversation.	Measuring a construct from a task that does not reliably elicit the target behavior.	Reflects **task–metric compatibility** and **construct validity** in Table 2 and **metric suggestions** embedded in Table 1.
Prioritize measures that are interpretable and useful.	Use scores or ratings that can guide treatment, reflect real-life communication, and be explained clearly to clients and families, such as informativeness, efficiency, or communicative success.	Reporting numbers that are difficult to interpret clinically or disconnected from treatment goals.	Reflects **interpretability** and **clinical utility** in Table 2.
Balance structure with functional relevance.	When possible, combine one structured task with one more functional task, for example, picture description plus conversation or narrative plus procedural discourse.	Assuming that a highly controlled task fully represents everyday communication.	Reflects **ecological relevance** in Table 2 and the **complementary strengths of elicitation types** in Table 1.
Leverage perceptual, semiautomated, and automated scoring methods to overcome time barriers.	Incorporate brief ratings, structured observations, or simple rubrics when transcription-heavy analysis is not possible. Use automated tools to reduce burden when available but verify outputs with clinical judgment.	Abandoning discourse assessment because comprehensive coding is too time consuming; assuming that automated output is accurate or sufficient on its own.	Reflects **clinical utility/feasibility** and **compatibility with automation and human judgment** in Table 2.
Document the context of performance.	Note cueing, fatigue, partner support, sensory issues, cultural/contextual fit, and other factors that shape performance.	Treating a sample as context-free or fully representative.	Reflects **interpretability**, **ecological relevance**, and **population/context fit** in Table 2.
Use within-participant methodology to track change consistently over time.	Leverage within-subject design methodologies (e.g., multiple baseline, ABAB, pre/post with random probes) to monitor progress while acknowledging that discourse associates with variability.	Single pre/post metrics; changing tasks or metrics across sessions.	Reflects **reliability** and **sensitivity to change** in Table 2.
Let practice inform future evidence.	Record which tasks and measures were feasible, meaningful, and sensitive to change in routine care.	Assuming that only formal research settings generate useful evidence.	Extends **replicability/ transparency** in Table 2 into practice-based evidence generation.

impose distinct cognitive–linguistic demands (Fergadiotis & Wright, 2011; B. C. Stark, 2019; B. C. Stark & Fukuyama, 2021; Ulatowska et al., 1981). For instance, narratives rely on temporal sequencing, causal coherence, and thematic elaboration, whereas procedural discourse emphasizes order, directive language, and clarity of instruction. These genre differences result in different language outputs—narrative tasks elicit more complex and propositionally rich speech, whereas procedural tasks yield simpler, shorter utterances. Importantly, these genre effects occur in both neurologically healthy speakers and those with communication disorders.

Psychometric evidence supports this view. Common discourse-derived variables show variable reliability depending on the elicitation task (e.g., B. C. Stark et al., 2023). Measures that appear stable across tasks may show poor test–retest reliability when examined within a single genre or sample length (Boyle, 2014; Brookshire & Nicholas, 1994). This undermines assumptions that discourse tasks are neutral containers for language behavior and underscores the need for genre- and task-specific validation. Without task-informed interpretation, discourse measures risk producing misleading or uninterpretable results.

Where We've Been and Where We Need to Go

A major concern is that discourse tasks and metrics are often selected based on tradition or convenience rather than empirical validation. Many commonly used tasks and metrics lack or do not report evidence for essential psychometric properties, such as reliability (across time, raters, or contexts), validity (in capturing the intended constructs for specific populations), and sensitivity to change (in detecting treatment-related improvements; Pritchard et al., 2018; B. C. Stark & Dalton, 2024). These limitations constrain their usefulness in clinical trials and translational research. If discourse is to function as a meaningful outcome in treatment studies, then both the elicitation tasks and the metrics used to analyze the resulting discourse must meet the same standards of measurement quality expected of other clinical instruments. The fact that a task or metric was not originally designed as an outcome measure does not automatically preclude its later use in that role. What matters is whether the specific task–metric pairing has demonstrated validity, reliability, and responsiveness for the intended purpose.

Construct validity depends not just on what a task measures but also on for whom it is appropriate. A picture description task optimized for aphasia may not reveal pragmatic impairments common in TBI or RHD. A mismatch between task demands and disorder profile compromises interpretability. For discourse assessment to function across conditions, task design must align with the cognitive–linguistic breakdowns specific to each population, and both tasks and metrics must be tested empirically in those populations (see Table 1). For example, a researcher studying narrative organization after TBI might select a video- or picture-based retell with macrostructural and coherence ratings rather than rely solely on a picture description optimized for lexical retrieval. A clinician working with a person with RHD whose primary goal is transactional success might combine a structured picture description for impairment characterization with a brief, personally relevant conversational or procedural task rated for informativeness and communicative effectiveness. These examples illustrate that the goal is construct validity: alignment of elicitation method and metric (outcome) with the population being asked.

This rigor extends to the metrics derived from those tasks. Although simple measures such as word counts are easy to compute and often reliable, they undeniably lack the depth needed to reflect functional communication. More functional metrics—such as informativeness—require clearly defined coding procedures and validation across contexts (Nicholas & Brookshire, 1993). A strong discourse task can be undermined by weak or mismatched metrics, just as a well-designed metric can underperform if applied to an inappropriate task.

To move the field forward, researchers must shift from assuming psychometric adequacy to actively demonstrating it. Only through deliberate, evidence-based design and evaluation can discourse assessment tools accurately capture meaningful changes in communication and support claims made in clinical and research contexts. Here, "discourse tools" refer broadly to the combined system of elicitation tasks, scoring procedures, and analytic metrics used to sample and interpret discourse. Indeed, emerging initiatives in discourse task design reflect this shift toward discourse tools that are theory driven, psychometrically evaluated, and tailored to specific populations or constructs. Examples include video-based narrative retell paradigms paired with semiautomated main concept analysis (Kurland et al., 2021) and efforts to examine the clinical utility of existing rating tools such as the Pragmatics Rating Scale (Iwashita & Sohlberg, 2019) and linguistic perceptual rating scales (Casilio et al., 2019). Some initiatives integrate discourse elicitation into treatment itself (Cannizzaro & Coelho, 2002; Dipper et al., 2024), enabling clinicians to measure change over time with tools that align to therapy targets.

Although researchers may have greater access to time, personnel, and analytic tools, psychometric standards do not stop at the research setting. Single-subject design is a cornerstone of clinical implementation because it allows clinicians to evaluate whether change is occurring within the individual, rather than assuming that any difference across sessions reflects treatment benefit. In practice, this might mean repeatedly sampling the same discourse goal—such as topic maintenance in conversation after TBI, informativeness during a procedural explanation, or narrative organization in a story retell—using a consistent elicitation genre and the same small set of interpretable outcome measures across sessions. Clinicians need not become full-time discourse researchers to do this well; rather, they can draw on the research methods training acquired in graduate school and through continuing education to ask psychometrically important questions in everyday practice: Am I using the same construct each time? Is this change larger than the variability I usually see in this client's discourse? Could improvement reflect familiarity with the task rather than treatment? Framed this way, clinical discourse assessment becomes not just descriptive but also evaluative, helping clinicians determine with greater confidence whether they are observing meaningful treatment-related change rather than the inherent fluctuation of discourse performance.

Another clear future direction is the partnership between theoretical and psychometric rigor with participatory design. Participatory design ensures that discourse tasks are not only theoretically and psychometrically grounded but also relevant, motivating, and feasible. Involving individuals with lived experience as well as clinicians helps define practical constraints, acceptable formats,

and ecologically valid prompts. Participatory tools—such as photo diaries, story grids, and visual rating cards—can facilitate engagement for individuals with language impairments (Wilson et al., 2015). Tasks and metrics developed through this process will likely elicit richer, more meaningful discourse and better align with the communicative contexts, priorities, and lived experiences of the people who use them. This principle is visible in aphasia intervention work such as Language Underpins Narrative in Aphasia, where treatment is personalized around the individual's own communication priorities and everyday language contexts, illustrating how discourse tasks and outcomes can be shaped not only by theory but also by what people with aphasia and clinicians judge to be meaningful and usable in daily life (Cruice et al., 2022).

There is also a clear future direction in improving how discourse is measured, not only to increase efficiency and reliability but also to better align elicitation tasks and metrics with specific clinical and research purposes. Importantly, the need for new tasks and metrics is not driven solely by time burden; rather, it reflects the need for better fit among task, population, construct, and purpose. At the same time, newer elicitation tasks can be designed to be briefer and more targeted, and newer metrics can emphasize interpretable, lower burden approaches such as perceptual ratings, semiautomated analyses, and clearly defined outcome measures that do not require exhaustive transcription or coding.

Where once a lack of tools and time hindered discourse analysis in clinical practice (Bryant et al., 2017; Cruice et al., 2020; B. C. Stark et al., 2021), technological advances in natural language processing, machine learning, and large language models (LLMs) offer exciting new capabilities for scaling discourse analysis. Automated pipelines can now extract lexical, syntactic, and discourse-level features rapidly and consistently, enabling analysis across large data sets and time points (MacWhinney, 2007; Themistocleous & Stark, 2026). This creates possibilities for longitudinal tracking, within-person change detection, normative modeling, and cross-diagnostic comparison. However, automation brings new responsibilities. It does not remove the need for careful design or human input; rather, it raises the stakes. Automated analysis amplifies the strengths and weaknesses of both task and metric. If a task fails to elicit meaningful discourse or a metric lacks interpretive value, then automated tools will produce misleading results, regardless of sophistication. For example, LLMs may overestimate global coherence (the extent to which discourse stays meaningfully related to the overall topic or main idea across the full sample) when language is grammatically fluent and superficially well connected, despite being semantically thin, repetitive, or pragmatically uninformative. At the same time, interest has also grown in validating perceptual rating scales as a more efficient but still clinically grounded approach to discourse assessment, complementing fully automated methods (Casilio et al., 2019; Iwashita & Sohlberg, 2019).

Whether perceptual or automated analyses are being used, very short samples and very verbose but semantically empty samples are challenging for both human raters and automated systems, although for different reasons. Very short samples may not provide enough evidence to support stable judgment of higher level discourse features such as informativeness, coherence, macrostructure, or pragmatic success. For example, a speaker with nonfluent aphasia may produce only a few words or short phrases in response to a narrative prompt; those utterances may still reveal important lexical or syntactic information, but they may be too sparse to judge whether the speaker can maintain a narrative theme, convey key story elements, or achieve communicative success. At the other extreme, very verbose but semantically thin discourse in a person with TBI can create a different interpretive problem. A speaker may produce many fluent utterances, complete sentences, and plausible local connections yet still fail to convey the essential propositions of a story, maintain global organization, or provide pragmatically relevant information. In such cases, both human raters and automated systems may be influenced by surface fluency, while the real difficulty lies in evaluating proposition-level content, macrostructural organization, and pragmatic adequacy. These challenges underscore that automation cannot replace trained clinical judgment. Human rating is not the gold standard simply by virtue of being human, and automated scoring is not inherently objective simply by virtue of being computational. Both are measurement systems, and both therefore need psychometric scrutiny: clear construct definition, adequate sampling, transparent scoring rules, reliability, validity, and interpretability.

Can and Should We Still Use Canonical Tasks?

Canonical discourse tasks (e.g., Cookie Theft, Cinderella) should continue to be used when they remain *fit for purpose*—that is, when their structure, available comparative data, and established analytic conventions support the specific clinical or research question better than available alternatives do. In particular, they remain especially useful when historical comparability matters, such as linking new findings to prior aphasia, TBI, dementia, or RHD literature or to shared data sets; when cross-study or cross-site consistency is a priority, as in multi-lab work, registries, or archival comparisons; when prior task-specific data are needed to inform study planning, including effect-size estimation and power analyses; when the task aligns well with the construct of interest, for example, when a structured picture description

Table 4. Scenarios for when canonical and noncanonical tasks may be appropriate, as well as considerations that accompany those scenarios.

Scenario	Purpose-driven task choice	Why this choice fits the goal	Key trade-off/implication
Research using a canonical task	Use **Cinderella** in a large TBI clinical trial.	Historical comparability with prior TBIBank (Elbourn et al., 2023) studies and shared analytic conventions are central to the study goal.	May be less sensitive to discourse difficulties that emerge when speakers must process unfamiliar material; may need a complementary task if study aims extend beyond what the canonical stimulus captures.
Research developing a novel task	Use a **video-based narrative retell** to study how adults with concussion summarize novel, dynamic events.	Captures discourse under conditions of new information processing rather than familiarity with a well-known story.	Requires stronger psychometric groundwork: clearly define the construct, determine whether parallel videos are needed, specify scoring procedures, evaluate interrater/test–retest reliability, ensure interpretability for the target population, and consider feasibility for eventual clinical translation.
Clinical use of a canonical task	Use a **canonical picture description task** with an adult with chronic aphasia after left-hemisphere stroke.	Provides a structured baseline sample of naming, sentence formulation, and informativeness with low memory burden and familiar analytic options such as core lexicon and main concept analysis.	May be less personally salient, motivating, and representative of everyday communication needs than a more individualized or functional discourse task; requires clinically disciplined measurement through a consistent task format, clearly defined outcomes, and repeated within-person sampling over time.
Clinical use of a noncanonical task	Acquire a **co-constructed** or **semistructured conversation** sample about a real workplace task with an adult with right-hemisphere disorder.	Supports evaluation of everyday communicative effectiveness in relation to the client's therapy goal rather than performance on a decontextualized prompt.	Less standardized and more vulnerable to partner, topic, or context effects but often more salient and functionally informative; requires clinically disciplined measurement through a consistent task format, clearly defined outcomes, and repeated within-person sampling over time.

Note. Note that coupling a canonical task with a noncanonical task can often be a wise decision and fit one's needs. TBI = traumatic brain injury.

is appropriate for examining lexical retrieval or informativeness; when the target population can engage with the task meaningfully without major cultural, experiential, sensory, or cognitive mismatch; and when the selected metric has empirical support for that task in that population. Table 4 provides research and clinical example scenarios illustrating when a canonical task may be appropriate, when a noncanonical alternative may be preferable, and how canonical tasks can also complement newer or more individualized discourse tasks.

A Call to Action

Ultimately, discourse tools are strongest when they are fit for purpose: when elicitation goals, sampling procedures,

Table 5. Call-to-action agenda items to drive the field of neurogenic/cognitive-communication discourse assessment and analysis forward.

Objective	Description
1. Define the purpose and select a fit-for-purpose task.	Choose a discourse task based on what you want to understand, support, or measure; the person or population of interest; and the practical realities of the setting. Canonical tasks remain useful when historical comparability, shared data sets, structured elicitation, or reduced task demands are central to the purpose; noncanonical tasks may be preferable when everyday relevance, novelty, or individual salience is more important.
2. Engage end users/people with lived experience.	Include clinicians and people with lived experience so that discourse tasks, supports, and outcomes are relevant, usable, and meaningful.
3. Choose meaningful and feasible measures.	Use metrics and scoring procedures that fit the task, reflect the construct of interest, are interpretable, and are sensitive to meaningful differences or change, while remaining feasible for the available time and resources.
4. Prioritize quality and consistency.	Use administration, scoring, and analysis procedures that are as reliable, valid, transparent, and repeatable as possible within the setting.
5. Use efficient tools thoughtfully.	Use automation, perceptual ratings, and other psychometrically robust tools to reduce scoring burden when appropriate. These tools can improve practicality, but they should still be matched to the construct of interest and supported by clear scoring procedures and clinical judgment.
6. Be clear about limitations.	State what the task and resulting measures can and cannot show and interpret findings in light of those boundaries.

Note. Ideally, the order of these should help design a research study and streamline clinical decision making.

and metrics are deliberately aligned with the construct of interest and the decisions they are meant to support. Table 5 translates this argument into a practical call to action for researchers and clinicians by outlining key priorities for developing discourse assessment that is psychometrically rigorous, clinically useful, and meaningful for assessment, treatment planning, and measuring change.

Data Availability Statement

TalkBank clinical banks (e.g., AphasiaBank, TBIBank) can be accessed by reading the rules and procedures at www.talkbank.org.

Acknowledgments

This study was supported by National Institute on Deafness and Other Communication Disorders Grant R01DC008524 (awarded to Brielle C. Stark, Co-Investigator, and Brian MacWhinney, Principal Investigator).

References

Bartels-Tobin, L. R., & Hinckley, J. J. (2005). Cognition and discourse production in right hemisphere disorder. *Journal of Neurolinguistics, 18*(6), 461–477. https://doi.org/10.1016/j.jneuroling.2005.04.001

Berube, S. K., Goldberg, E., Sheppard, S. M., Durfee, A. Z., Ubellacker, D., Walker, A., Stein, C. M., & Hillis, A. E. (2022). An analysis of right hemisphere stroke discourse in the Modern Cookie Theft picture. *American Journal of Speech-Language Pathology, 31*(5S), 2301–2312. https://doi.org/10.1044/2022_AJSLP-21-00294

Berube, S., Nonnemacher, J., Demsky, C., Glenn, S., Saxena, S., Wright, A., Tippett, D. C., & Hillis, A. E. (2019). Stealing cookies in the twenty-first century: Measures of spoken narrative in healthy versus speakers with aphasia. *American Journal of Speech-Language Pathology, 28*(1S), 321–329. https://doi.org/10.1044/2018_AJSLP-17-0131

Boyle, M. (2014). Test–retest stability of word retrieval in aphasic discourse. *Journal of Speech, Language, and Hearing Research, 57*(3), 966–978. https://doi.org/10.1044/2014_JSLHR-L-13-0171

Brookshire, R. H., & Nicholas, L. E. (1994). Speech sample size and test–retest stability of connected speech measures for adults with aphasia. *Journal of Speech and Hearing Research, 37*(2), 399–407. https://doi.org/10.1044/jshr.3702.399

Bryant, L., Spencer, E., & Ferguson, A. (2017). Clinical use of linguistic discourse analysis for the assessment of language in aphasia. *Aphasiology, 31*(10), 1105–1126. https://doi.org/10.1080/02687038.2016.1239013

Cannizzaro, M. S., & Coelho, C. A. (2002). Treatment of story grammar following traumatic brain injury: A pilot study. *Brain Injury, 16*(12), 1065–1073. https://doi.org/10.1080/02699050210155230

Casilio, M., Rising, K., Beeson, P. M., Bunton, K., & Wilson, S. M. (2019). Auditory-perceptual rating of connected speech in aphasia. *American Journal of Speech-Language Pathology, 28*(2), 550–568. https://doi.org/10.1044/2018_AJSLP-18-0192

Cruice, M., Aujla, S., Bannister, J., Botting, N., Boyle, M., Charles, N., Dhaliwal, V., Grobler, S., Hersh, D., Marshall, J., Morris, S., Pritchard, M., Scarth, L., Talbot, R., & Dipper, L. (2022). Creating a novel approach to discourse treatment through coproduction with people with aphasia and speech and language therapists. *Aphasiology, 36*(10), 1159–1181. https://doi.org/10.1080/02687038.2021.1942775

Cruice, M., Botting, N., Marshall, J., Boyle, M., Hersh, D., Pritchard, M., & Dipper, L. (2020). UK speech and language therapists' views and reported practices of discourse analysis in aphasia rehabilitation. *International Journal of Language & Communication Disorders, 55*(3), 417–442. https://doi.org/10.1111/1460-6984.12528

Dalton, S. G., AL Harbi, M., Berube, S., & Hubbard, H. I. (2024). Development of main concept and core lexicon checklists for the original and modern Cookie Theft stimuli. *Aphasiology, 38*(12), 1975–1999. https://doi.org/10.1080/02687038.2024.2340794

Dipper, L., Devane, N., Barnard, R., Botting, N., Boyle, M., Cockayne, L., Hersh, D., Magdalani, C., Marshall, J., Swinburn, K., & Cruice, M. (2024). A feasibility randomised waitlist-controlled trial of a personalised multi-level language treatment for people with aphasia: The remote LUNA study. *PLOS ONE, 19*(6), Article e0304385. https://doi.org/10.1371/journal.pone.0304385

Dutta, M., Murray, L. L., Park, H., Burklow, E., Bose, A., Kim, H., Greenslade, K., Ramage, A. E., Combs, C., Balasubramanian, A., & Casilio, M. (2025). Let's chat about spoken discourse: A tutorial to support use of spoken discourse analysis when providing aphasia clinical services. *American Journal of Speech-Language Pathology, 34*(4), 1931–1966. https://doi.org/10.1044/2025_AJSLP-24-00534

Elbourn, E., MacWhinney, B., Fromm, D., Power, E., Steel, J., & Togher, L. (2023). TBIBank: An international shared database to enhance research, teaching and automated language analysis for traumatic brain injury populations. *Archives of Physical Medicine and Rehabilitation, 104*(5), 824–829. https://doi.org/10.1016/j.apmr.2022.12.192

Fergadiotis, G., & Wright, H. H. (2011). Lexical diversity for adults with and without aphasia across discourse elicitation tasks. *Aphasiology, 25*(11), 1414–1430. https://doi.org/10.1080/02687038.2011.603898

Goodglass, H., & Kaplan, E. (1972). *Boston Diagnostic Aphasia Examination*. Lea & Febiger.

Greenslade, K. J., Honan, C., Harrington, L., Kenealy, L., Ramage, A. E., & Bogart, E. (2024). Wishes, beliefs, and jealousy: Use of mental state terms in *Cinderella* retells after traumatic brain injury. *Frontiers in Human Neuroscience, 18*. https://doi.org/10.3389/fnhum.2024.1386227

Greenslade, K. J., Stuart, J. E. B., Richardson, J. D., Dalton, S. G., & Ramage, A. E. (2020). Macrostructural analyses of Cinderella narratives in a large nonclinical sample. *American Journal of Speech-Language Pathology, 29*(4), 1923–1936. https://doi.org/10.1044/2020_ajslp-19-00151

Grimes, N. (2005). *Walt Disney's Cinderella*. Random House.

Hadar, U., Jones, C., & Mate-Kole, C. (1987). The disconnection in anomic aphasia between semantic and phonological lexicons. *Cortex, 23*(3), 505–517. https://doi.org/10.1016/S0010-9452(87)80011-4

Hadar, U., Ticehurst, S., & Wade, J. P. (1991). Crossed anomic aphasia: Mild naming deficits following right brain damage in a dextral patient. *Cortex, 27*(3), 459–468. https://doi.org/10.1016/S0010-9452(13)80042-1

Hux, K., & Frodsham, K. (2023). Speech and language characteristics of neurologically healthy adults when describing the Modern Cookie Theft picture: Mixing the new with the old. *American Journal of Speech-Language Pathology, 32*(3), 1110–1130. https://doi.org/10.1044/2022_AJSLP-22-00291

Iwashita, H., & Sohlberg, M. M. (2019). Measuring conversations after acquired brain injury in 30 minutes or less: A comparison of two pragmatic rating scales. *Brain Injury, 33*(9), 1219–1233. https://doi.org/10.1080/02699052.2019.1631487

Kurland, J., Liu, A., & Stokes, P. (2021). Phase I test development for a brief assessment of transactional success in aphasia: Methods and preliminary findings of main concepts in non-aphasic participants. *Aphasiology, 37*(1), 39–68. https://doi.org/10.1080/02687038.2021.1988046

Leaman, M. C., & Archer, B. (2023). Choosing discourse types that align with person-centered goals in aphasia rehabilitation: A clinical tutorial. *Perspectives of the ASHA Special Interest Groups, 8*(2), 254–273. https://doi.org/10.1044/2023_PERSP-22-00160

MacWhinney, B. (2007). The TalkBank project. In J. C. Beal, K. P. Corrigan, & H. L. Moisl (Eds.), *Creating and digitizing language corpora, Volume 1: Synchronic databases* (pp. 163–180). Palgrave Macmillan. https://doi.org/10.1057/9780230223936_7

Nicholas, L. E., & Brookshire, R. H. (1993). A system for quantifying the informativeness and efficiency of the connected speech of adults with aphasia. *Journal of Speech and Hearing Research, 36*(2), 338–350. https://doi.org/10.1044/jshr.3602.338

Pritchard, M., Hilari, K., Cocks, N., & Dipper, L. (2018). Psychometric properties of discourse measures in aphasia: Acceptability, reliability, and validity. *International Journal of Language & Communication Disorders, 53*(6), 1078–1093. https://doi.org/10.1111/1460-6984.12420

Stark, B. C. (2019). A comparison of three discourse elicitation methods in aphasia and age-matched adults: Implications for language assessment and outcome. *American Journal of Speech-Language Pathology, 28*(3), 1067–1083. https://doi.org/10.1044/2019_AJSLP-18-0265

Stark, B. C. (2023). Historical review of research in discourse deficits and its recent advancement. In A. P.-H. Kong (Ed.), *Spoken discourse impairments in the neurogenic populations: A state-of-the-art, contemporary approach* (pp. 3–22). Springer. https://doi.org/10.1007/978-3-031-45190-4_1

Stark, B. C., Alexander, J. M., Hittson, A., Doub, A., Igleheart, M., Streander, T., & Jewell, E. (2023). Test–retest reliability of microlinguistic information derived from spoken discourse in persons with chronic aphasia. *Journal of Speech, Language, and Hearing Research, 66*(7), 2316–2345. https://doi.org/10.1044/2023_JSLHR-22-00266

Stark, B. C., Bryant, L., Themistocleous, C., den Ouden, D.-B., & Roberts, A. C. (2022). Best practice guidelines for reporting spoken discourse in aphasia and neurogenic communication disorders. *Aphasiology, 37*(5), 761–784. https://doi.org/10.1080/02687038.2022.2039372

Stark, B. C., & Dalton, S. G. (2024). A scoping review of transcription-less practices for analysis of aphasic discourse and implications for future research. *International Journal of Language & Communication Disorders, 59*(5), 1734–1762. https://doi.org/10.1111/1460-6984.13028

Stark, B. C., Dutta, M., Murray, L. L., Fromm, D., Bryant, L., Harmon, T. G., Ramage, A. E., & Roberts, A. C. (2021). Spoken discourse assessment and analysis in aphasia: An international survey of current practices. *Journal of Speech, Language, and Hearing Research, 64*(11), 4366–4389. https://doi.org/10.1044/2021_JSLHR-20-00708

Stark, B. C., & Fukuyama, J. (2021). Leveraging big data to understand the interaction of task and language during monologic spoken discourse in speakers with and without aphasia. *Language, Cognition and Neuroscience, 36*(5), 562–585. https://doi.org/10.1080/23273798.2020.1862258

Stark, J. A. (2010). Content analysis of the fairy tale *Cinderella*—A longitudinal single-case study of narrative production: "From rags to riches." *Aphasiology, 24*(6–8), 709–724. https://doi.org/10.1080/02687030903524729

Stockbridge, M. D., & Newman, R. (2019). Enduring cognitive and linguistic deficits in individuals with a history of concussion. *American Journal of Speech-Language Pathology, 28*(4), 1554–1570. https://doi.org/10.1044/2019_AJSLP-18-0196

Themistocleous, C., & Stark, B. C. (2026). Language biomarker screening using AI: A transdiagnostic approach to the brain. *Scientific Reports, 16*(1), Article 4207. https://doi.org/10.1038/s41598-025-34257-z

Ulatowska, H. K., North, A. J., & Macaluso-Haynes, S. (1981). Production of narrative and procedural discourse in aphasia. *Brain & Language, 13*(2), 345–371. https://doi.org/10.1016/0093-934X(81)90100-0

Wilson, S., Roper, A., Marshall, J., Galliers, J., Devane, N., Booth, T., & Woolf, C. (2015). Codesign for people with aphasia through tangible design languages. *CoDesign, 11*(1), 21–34. https://doi.org/10.1080/15710882.2014.997744