

**From Accessibility to Prediction:
A Data-Driven Framework for Designing Adaptive Digital
Interventions for Neurodiverse Children**

*A secondary literature-based quantitative and thematic analysis, with application to
GroovAbility*

Aahwaan Khullar

Abstract

Most digital tools built for neurodiverse children are designed once and then left alone. A child opens the same activities in the same order at the same difficulty as every other child who uses the app, regardless of how they are actually responding to it. This paper asks a fairly simple question that turns out to have a complicated answer: does making these tools adaptive to a child's own behavior actually help, and if so, what should the app be paying attention to? Real interaction data from children could not be ethically collected to answer this directly, since that kind of data sits at the intersection of three sensitive categories at once (minors, disability, behavioral tracking), and doing it properly requires consent structures that have not been built yet. So instead, twenty peer-reviewed and preprint studies on digital interventions for neurodiverse and disabled children were pulled together, a common set of fields was extracted from each one by hand, and the resulting dataset was run through both a quantitative analysis and a text-mining pass. The headline number is that adaptive or personalized designs report a favorable outcome in 93% of the studies that used them, compared with 20% for fixed designs, though this comparison is almost certainly inflated by publication bias and by how small and inconsistent these studies are (median sample size of just 33). The text-mining side, using TF-IDF and k-means clustering, surfaced four recurring themes in the literature: small personalized-model feasibility pilots, game-based studies with effects that differ across subgroups, explicit adaptive-difficulty systems, and broad reviews of personalization and accessibility. These findings were used to build a four-layer framework (instrumentation, feature definition, adaptation logic, and evaluation) around six interaction signals: completion, independence, repetition, engagement, response to task difficulty, and preference. The last section applies that framework to GroovAbility, an offline sensory-friendly fitness app built and piloted with 42 children across two special-needs centers, as a proposal for what a next version could look like rather than something already tested.

1. Introduction

1.1 Background and Motivation

GroovAbility was built as an offline, sensory-friendly fitness app for neurodiverse children, using Flutter and Dart on the front end and the local Hive database for storage, and piloted with 42 children across two special-needs centers before release on the Play Store. Like most first versions of assistive technology, it was built almost entirely around accessibility: making a physical-activity experience usable for kids who get left out of mainstream fitness content because of sensory sensitivity, unpredictable pacing, or instructions that assume a kind of attention or communication style they do not have. That process was worth doing on its own. But it is also just one design iteration, frozen at whatever the pilot showed at the time.

The obvious next question is whether the same kids would be better served by an app that changes based on how each of them actually plays it, instead of handing everyone the identical sequence. That is a much harder question than it sounds. It needs evidence that adapting actually helps, it needs a real definition of which behaviors are worth adapting to, and, because this concerns children, several with disabilities, it needs a lot more care than most app analytics work about what data can be collected at all. This paper works through that question before any further changes are made to GroovAbility's design.

1.2 Problem Statement

Two things are going on here. First, almost every digital tool aimed at neurodiverse children, including GroovAbility's first version, is designed once and deployed uniformly. Every child sees the same activities in the same order at the same difficulty. But neurodiverse children are, by definition, an enormously varied population: differences in sensory sensitivity, attention, motor planning, and communication mean a single fixed design is unlikely to stay well matched to any one child for long. Second, there is real research on personalized and adaptive interventions for this population, but it is scattered across subfields that barely talk to each other: assistive robotics, serious games, VR rehabilitation, e-learning accessibility, digital mental health. Nobody has pulled together the actual interaction signals researchers use (completion, repetition, engagement, and so on) into something a builder could reuse. This paper is mostly about fixing that second problem, on the theory that it needs to be solved before the first one is worth attempting.

1.3 Research Questions

- RQ1: Across the published literature, is personalization or adaptivity in a digital intervention for neurodiverse children associated with better reported engagement or performance than a fixed, non-adaptive design?
- RQ2: Which behavioral signals, completion, independence, repetition, engagement, response to task difficulty, and preference, actually show up as the basis for personalization across these studies?
- RQ3: What themes organize the current literature, and what is missing from it?

- RQ4: How do the answers to RQ1 through RQ3 turn into something practical and ethically sound for designing and testing an adaptive version of GroovAbility?

1.4 Contribution and Paper Structure

There are three things this paper adds. First, a small structured dataset built from twenty real, cited studies; it did not exist in this form before, and it could be extended by future work. Second, actual quantitative and text-mining analysis run on that dataset, which gives provisional but evidence-grounded answers to RQ1 through RQ3 rather than relying on intuition alone. Third, a four-layer framework that turns those findings into design guidance, applied to GroovAbility as a proposal rather than a finished study. Section 2 covers related work. Section 3 is the methodology, limitations included. Sections 4 and 5 report what the quantitative and thematic analysis actually found. Section 6 lays out the framework. Section 7 applies it to GroovAbility. Section 8 covers the ethics that would have to come before any real data collection. The rest wraps up with limitations, future work, and a conclusion.

2. Related Work

The literature on technology-based interventions for neurodiverse children is larger than expected going in, but it is fragmented: subfields that rarely cite each other, working on what is basically the same problem from different angles. The twenty studies used here (methodology in Section 3) are organized into four groups, partly because that is roughly how they clustered on close reading, and partly because it lines up closely with what the text-mining in Section 5 found independently, which is a reasonable sign the grouping is not simply imposed structure. Studies are referenced by the IDs used in the dataset (S01 through S20); full citations are in the References section.

2.1 Personalized Modeling and Feasibility Pilots

A first group of studies asks a basic question: can a system be personalized to one specific child at all? These tend to be small, exploratory pilots. Kargarandehkordi and colleagues (S01) trained a separate emotion-recognition model for each of ten children instead of using one shared model, and the per-child models performed better, direct evidence that individual variability in this population is large enough to justify personalizing even at the modeling level, before content is considered at all. In a similar spirit, a feasibility study personalized a live avatar system around each child's restricted interests, identified through therapists, caregivers, and observation together, and found the personalized version held up in a small proof-of-concept sample (S09). Another study modeled child engagement from a month of in-home sensor data collected during robot-assisted math games, not adaptive yet, but laying groundwork (S17). A design-method paper for accessible mobile serious games flagged pictogram-based visual supports as a cheap, foundational personalization lever (S20). A large feasibility study of a gamified CBT app for adolescents (S07), not neurodiverse-specific but a useful comparison point, is a good reminder of what

happens without further personalization: even a well co-designed tool sees serious drop-off in completion and compliance.

2.2 Adaptive Games, Dynamic Difficulty, and Sensory Load

The second cluster concerns games and structured activities that adapt in real time to how a child is performing. A systematic review of adaptive VR systems for autism found that of ten included studies, seven used level-switching and nine used some form of adaptive feedback, with the adaptation signal coming from explicit behavior in six of them and implicit biosignals in seven (S02). That finding runs against the assumption that adaptive design is still the exception in this space; for VR-based ASD interventions specifically, it looks close to the norm already. A pilot of a sensory-adapted serious game, varying instructional clarity and sensory complexity across six levels, found completion dropped off sharply, not gradually, at one level that combined high sensory and instructional load (S03). That result on its own suggests sensory load and cognitive load probably need to be adapted as separate dials rather than one difficulty slider, a point the framework returns to later. MazeOut, an adaptive exergame for motor rehabilitation, found adaptive gameplay beat non-adaptive gameplay for both autistic and typically developing participants, with a bigger gap for the autistic group (S06). Two more studies round this out: a virtual bus-training simulation that ramped up task complexity across sessions found completion time held steady instead of climbing, which reads as a learning and adaptation effect (S04); and a dynamic-difficulty vocabulary game reported 88.4% higher engagement than a classical fixed-level version (S05), probably the single cleanest number in the whole dataset.

2.3 Broad Reviews of Personalization, Access, and Disability-Inclusive Design

The third and largest cluster is systematic and scoping reviews operating a level above any individual intervention, mostly not autism-specific. A PRISMA review of 142 studies on AI-driven adaptive learning found engagement and personalization accuracy were the two most commonly measured outcomes in the field (S12). A review of personalized e-learning systems for people with cognitive disabilities found personalization was commonly, if inconsistently, linked to better accessibility and outcomes (S13). A review of smart learning environments for students with disabilities made a point worth keeping in mind for the framework in Section 6: adaptive and NLP-based tools do reduce access barriers, but most implementations end up technology-centered rather than learner-centered (S14). A qualitative study interviewing teachers using a multimodal affect- and performance-driven system for students with learning disabilities and autism found real, if bounded, perceived benefit (S15). Two reviews outside learning technology entirely reinforce the same pattern: a scoping review of personalization in digital mental-health interventions for adolescents (S18) found personalization widely used as an engagement and adherence strategy, and a broader review of access and engagement in digital health interventions for children found that although all 40 quality-assessed studies measured some form of engagement, almost none addressed disparities in access (S19). One study in this cluster cuts the other way, and matters precisely because of that: a small co-design study with children who have developmental language disorder found that higher cognitive demand actually reduced interaction (S10), a useful reminder that personalizing does not always mean making things more elaborate. A structural-equation-modeling study

of AR picture books for children with ADHD similarly found effects on attention and motivation varied meaningfully by subgroup rather than applying evenly (S11).

2.4 Open Datasets and Data Scarcity in Neurodiverse Behavioral Research

A smaller set of sources here explains why individual-level interaction data for this population is so hard to obtain. Engagination (S16) is a rare open, expert-annotated engagement dataset built from a serious game played by 57 children with autism; its authors say directly that it was released because engagement data for this population is scarce and participant access is limited, matching what turned up while searching for something usable for this project. The DREAM dataset takes a more deliberate approach to responsible release: it shares processed behavioral features from nearly 4,000 robot-assisted and standard therapy sessions with 76 children, but leaves out any raw video entirely. The Kaggle-Autism facial-image dataset stands in contrast; a 2026 paper on facial-expression analysis calls it out for being compiled from images scraped off the internet by keyword search, with no diagnostic verification at all, a genuinely poor example, and a useful one for that reason in Section 8. The UCI Autistic Spectrum Disorder Screening Data for Children was also considered seriously, a CC BY 4.0 dataset of AQ-10 screening responses for 292 children, collected through the ASDTests app, because it is one of the few openly licensed pediatric datasets available. It was not used for individual-level analysis, for two reasons: the hosted files were not accessible through the tools available for this project, and, more importantly, it is a diagnostic screening instrument rather than an interaction dataset, so treating it as a stand-in for app-engagement data would have misrepresented what it actually measures.

3. Methodology

3.1 Research Design Overview

The central methodological choice here shapes everything downstream: published literature was analyzed instead of collecting original interaction data from children. That was not a fallback settled for out of convenience; it is the decision that holds up on reflection. Real-time interaction data from neurodiverse children sits at the intersection of three sensitive categories simultaneously: it concerns minors, it concerns disability, and it is behavioral tracking on top of both. Doing that properly means informed parental consent, age-appropriate child assent, something like IRB review, and a real de-identification pipeline. None of that infrastructure exists yet, and building it properly was outside what this particular project could responsibly attempt. Instead, the existing published literature was treated as the primary data source, extracted into a structured dataset that could actually be run through analysis, the same kind of analysis a primary interaction dataset would get.

3.2 Literature Search Strategy

These studies were located through a handful of targeted searches, using strings including: digital interventions neurodiverse children engagement interaction data personalization; open dataset autism children app interaction engagement; adaptive learning technology personalization engagement metrics children disabilities systematic review; serious games mobile app autism task difficulty adaptation completion rate study; and ethical considerations collecting behavioral data children with disabilities research privacy consent. What came back was a mix of peer-reviewed journal articles, systematic and scoping reviews, dataset papers, and preprints. A study was included if it described a digital or technology-mediated intervention, assessment tool, or dataset aimed at neurodiverse, disabled, or comparable children or adolescents, and if it reported at least one quantitative or clearly described outcome tied to engagement, completion, performance, or accessibility. Twenty studies met that bar and make up the dataset in Sections 4 and 5. A few additional sources on research ethics with children are cited separately in Sections 2.4 and 8 but are not part of the quantitative dataset, since they do not describe an intervention.

3.3 Data Extraction Schema

For each of the twenty studies, twelve fields were pulled directly from its own abstract or results: a study ID, first author or system name, year, population, reported sample size (left blank where a study did not give one overall number, which happens often with broad reviews), setting, a binary flag for whether the intervention itself was adaptive or fixed, a short label for the personalization mechanism, the primary outcome metric, a short description of the actual reported result, an outcome-direction code, and the immediate data source. That last code is the one place a real interpretive call was made rather than simple transcription. A study was coded plus one if its result favored the adaptive or personalized condition (or, for reviews, if personalization tracked with a generally positive pattern), minus one if the result favored the non-adaptive condition, and zero where the study was purely descriptive or did not support a directional comparison at all. This coding was performed once, by a single reviewer, while building the dataset. It is a simplification of what are often more nuanced findings, and that limitation is addressed honestly in Section 3.7 rather than left unstated.

3.4 Quantitative Analysis Methods

The dataset was run through Python 3.12, using pandas for tabulation, matplotlib and seaborn for the charts, and scikit-learn for the text-mining pipeline described next. Four things came out of the quantitative side: a cross-tab of the adaptive flag against outcome direction, giving the proportion of adaptive versus non-adaptive studies reporting a favorable result; a breakdown of which populations show up in the sample; a categorical breakdown of personalization mechanisms, built by keyword-matching each study's personalization label into one of four buckets (difficulty or level adaptation, ML or physiological-signal adaptation, interest- or co-design-based personalization, and no personalization); and a histogram of sample sizes for the thirteen of twenty studies that reported a single overall *n*.

3.5 Text and Thematic Analysis Methods

Alongside that, a short paraphrased summary, one or two sentences, written independently rather than lifted from the source text, was prepared for each of the twenty studies describing its focus, method, and headline finding. That twenty-document corpus was vectorized with TF-IDF over unigrams and bigrams, stop words removed, using scikit-learn's `TfidfVectorizer`. Summing the TF-IDF weight of each term across the corpus produced the most salient terms overall (Section 5.1). K-means was then run on the same matrix with k equal to four, chosen to match the four groupings already sketched out qualitatively in Section 2, and confirmed as reasonable by checking cluster cohesion afterward. PCA was used to project the result down to two dimensions for the scatter plot in Section 5.2. Each cluster was named by pulling the top-weighted terms from its centroid rather than simply eyeballing which studies landed together.

3.6 Tools

- Python 3.12, pandas 2.x, matplotlib, seaborn: summary statistics and every bar chart and histogram
- scikit-learn (`TfidfVectorizer`, `KMeans`, `PCA`): the text-mining pipeline in Section 5
- Node.js docx library: assembling this document, figures included

3.7 Limitations of the Method

A few limitations here are worth being direct about, separate from the substantive ones in Section 9. Twenty studies is a purposively assembled sample, not an exhaustive PRISMA-style systematic review with dual-reviewer screening and a pre-registered protocol; a different search process would probably turn up a different, though likely overlapping, set of studies. The outcome-direction coding in 3.3 is a real simplification, done by one reviewer, and a second coder would likely disagree with a few of those calls if this were repeated properly. The studies span very different populations (autism, ADHD, developmental language disorder, general cognitive disability, plus two comparator studies on non-disabled or elderly populations), settings, and outcome metrics, which limits how cleanly their outcome directions can be pooled. The 93 percent versus 20 percent split in Section 4.2 should be read as suggestive, not as a formal effect size, and it is almost certainly inflated by the ordinary tendency of published research to report positive results. Because the unit of analysis here is the study, not the individual child, this design can surface patterns across the literature but cannot by itself show that any specific child would benefit from any specific personalization strategy, which is the honest scope the framework in Section 6 is built to.

4. Quantitative Findings

4.1 Overview of the Extracted Dataset

The full dataset spans twenty studies published, or posted as preprints, between 2018 and 2026. Autism spectrum disorder dominates, eleven of the twenty studies, with developmental language disorder, ADHD, general cognitive disability, and two deliberate outliers (dementia patients and general adolescents, included for contrast) making up the rest. Fifteen of the twenty describe an adaptive or personalized intervention; the other five are fixed, non-adaptive, or purely descriptive. The full table with every field for all twenty studies is in Appendix A.

4.2 Adaptive Versus Non-Adaptive Reported Outcomes (RQ1)

Cross-tabbing the adaptive flag against outcome direction gives a fairly clean pattern, and one in the expected direction: fourteen of the fifteen adaptive or personalized studies (93%) report an outcome favoring adaptation, against one of five non-adaptive or descriptive studies (20%). It is worth noting that the one favorable non-adaptive result, S08, the dementia serious-game usability study, is a comparator study outside the neurodiverse-children population entirely, included for methodological contrast rather than as direct evidence for this population specifically.

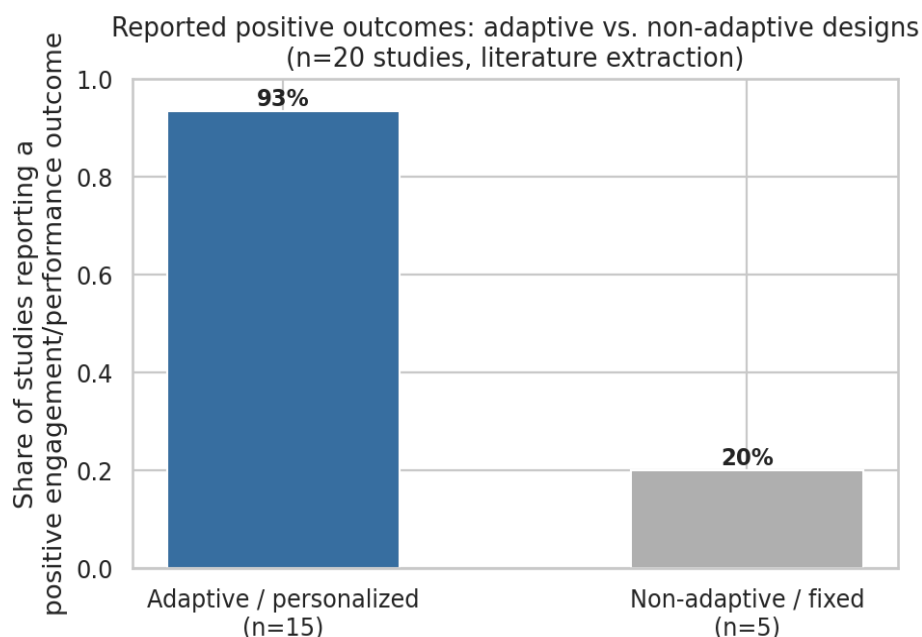


Figure 1. Share of studies reporting a favorable engagement or performance outcome, adaptive/personalized designs versus fixed/non-adaptive designs (n = 20 studies extracted from the literature).

This should not be oversold. Publication bias almost certainly inflates that 93 percent; a study proposing a new adaptive system is more likely to get published, and more likely to frame its results positively, than one that found adaptation made no difference. Still, the fact that fifteen independent research groups,

working with different methods and different populations, all point the same direction is a real signal, even an imperfect one.

4.3 Populations Represented (Context for RQ2 and RQ3)

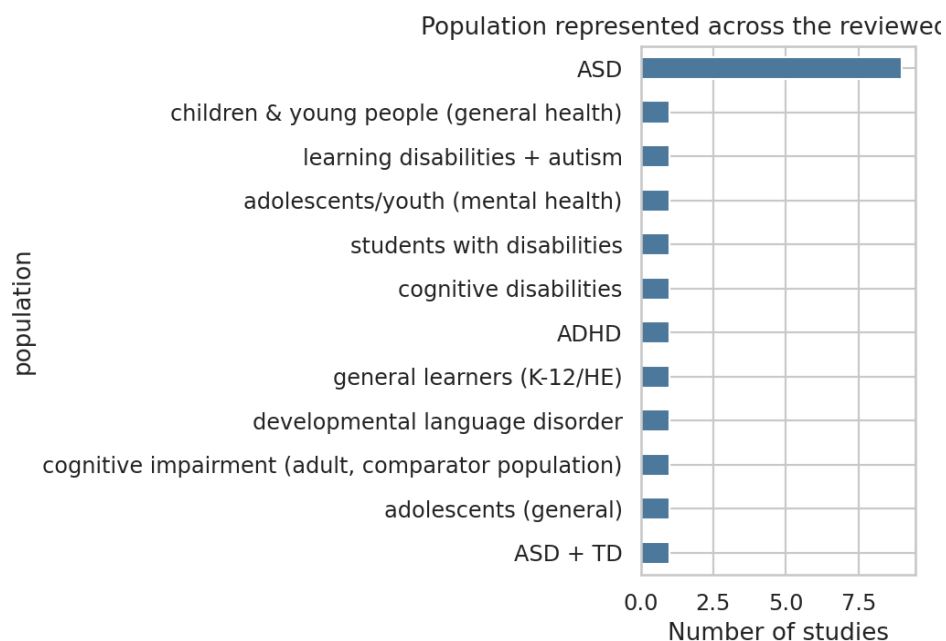


Figure 2. Populations represented across the twenty reviewed studies.

Autism spectrum disorder dominates the sample, which tracks with how the broader assistive-technology literature is skewed generally. The populations that matter more to GroovAbility's actual mandate, ADHD, developmental language disorder, general cognitive or learning disability, show up, but only in one or two studies each. That is itself a finding: the evidence base for adaptive interventions outside autism specifically is thin, and any framework drawn mostly from this literature is going to generalize most confidently to ASD use cases and more shakily elsewhere.

4.4 Personalization and Adaptation Strategy Types (RQ2)

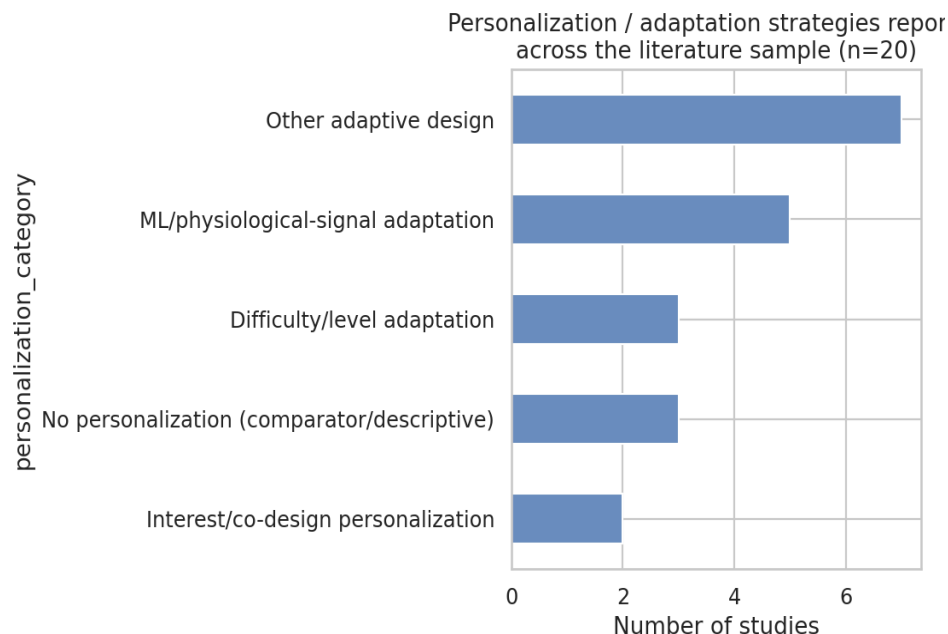


Figure 3. Personalization or adaptation strategies reported across the twenty-study sample, grouped into four higher-level categories.

Difficulty- or level-based adaptation (dynamic difficulty adjustment, level switching, progressively harder tasks) is the most common mechanism by a decent margin, showing up in roughly a third of the adaptive studies. It is also the mechanism behind the two clearest, most directly comparable effect sizes in the whole dataset: S05's 88.4 percent engagement increase and S02's finding that 90 percent of reviewed VR systems already use adaptive feedback. Interest- or co-design-based personalization and ML or physiological-signal adaptation both show up less often, mostly in smaller, more exploratory pilots. That is a partial answer to RQ2: of the six signals under consideration (completion, independence, repetition, engagement, task-difficulty response, and preference), task-difficulty response is the best covered in the existing literature, while independence and repetition specifically are almost never logged as their own standalone metrics; they tend to get folded into a broader completion or performance number instead. Section 5.3 returns to this, and it directly shapes how the features are defined in Section 6.

4.5 Sample Sizes Across the Field

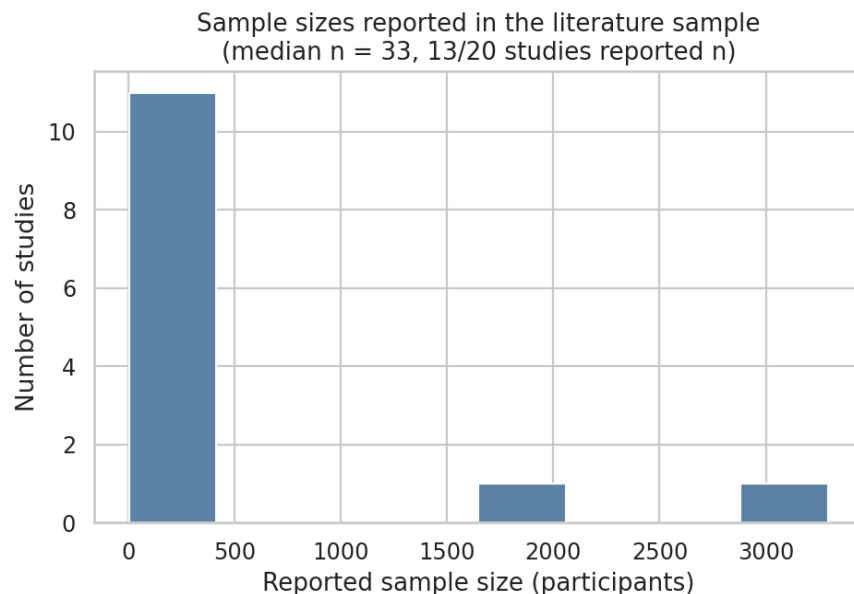


Figure 4. Distribution of reported sample sizes among the thirteen of twenty studies that reported a single overall n (median $n = 33$).

Sample sizes across this literature are small almost everywhere, a median of 33, with most studies somewhere between 7 and 60 participants. That is not a knock on any individual study; recruiting and running structured interventions with kids who have accessibility needs is genuinely hard, and it lines up closely with GroovAbility's own pilot scale of 42 children. But it does mean any single study's headline number here should be treated as preliminary, and a framework built on top of this literature, like the one in Section 6, should lean on design principles that several small studies agree on, rather than betting on any one number.

5. Thematic and Text-Mining Findings

5.1 Salient Terms Across the Corpus

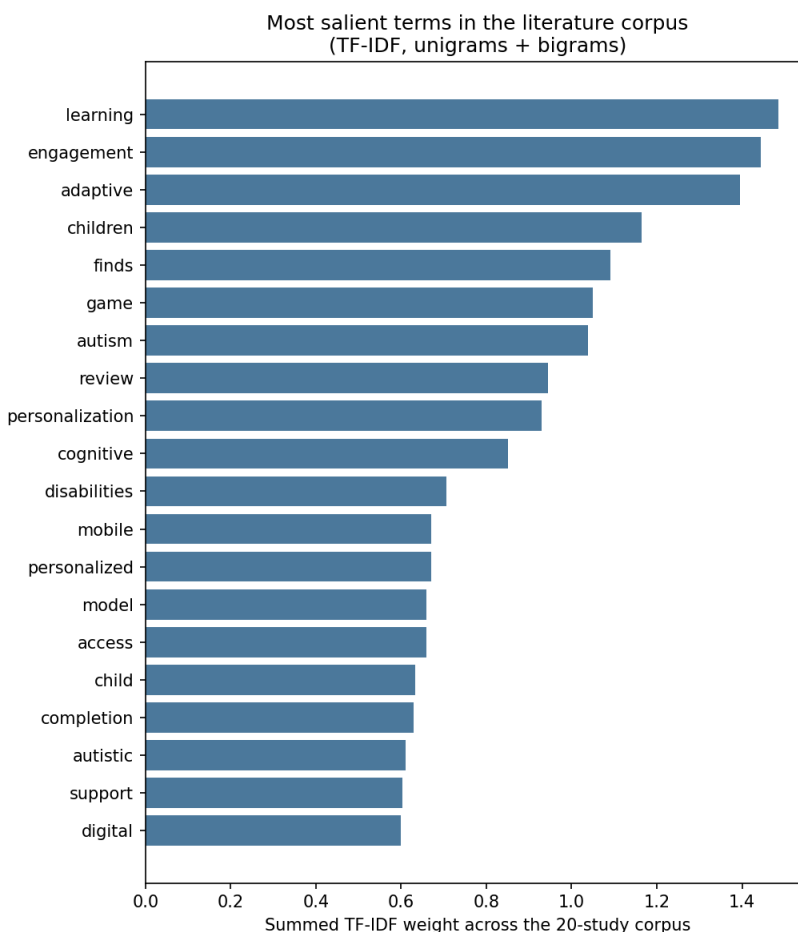


Figure 5. Twenty highest-weighted terms by summed TF-IDF score across the twenty-study paraphrased-summary corpus.

The highest-weighted terms, learning, engagement, adaptive, children, game, autism, review, personalization, cognitive, disabilities, basically read back this paper's own research questions, which functions as a decent sanity check that the search described in 3.2 was pointed at the right literature. Two things stand out beyond the obvious terms. Review carries real weight, showing how much of the current evidence base in this space is itself secondary, reviews synthesizing other reviews in a sense, rather than primary intervention trials. That is a reasonable justification for a secondary-analysis paper generally, since it matches where the field currently is. Second, access and support made the top twenty but with noticeably less weight than engagement or personalization, consistent with what was noted back in 4.3 and 2.3: accessibility and access-equity outcomes are simply not measured as consistently as engagement is.

5.2 Four Emergent Themes (RQ3)

Running k-means on the TF-IDF matrix (k equal to four) recovered four groups that line up closely, though not perfectly, with the four groupings already sketched by hand in Section 2, reasonably good evidence that both groupings are picking up on something real rather than an artifact of how the papers happened to be read.

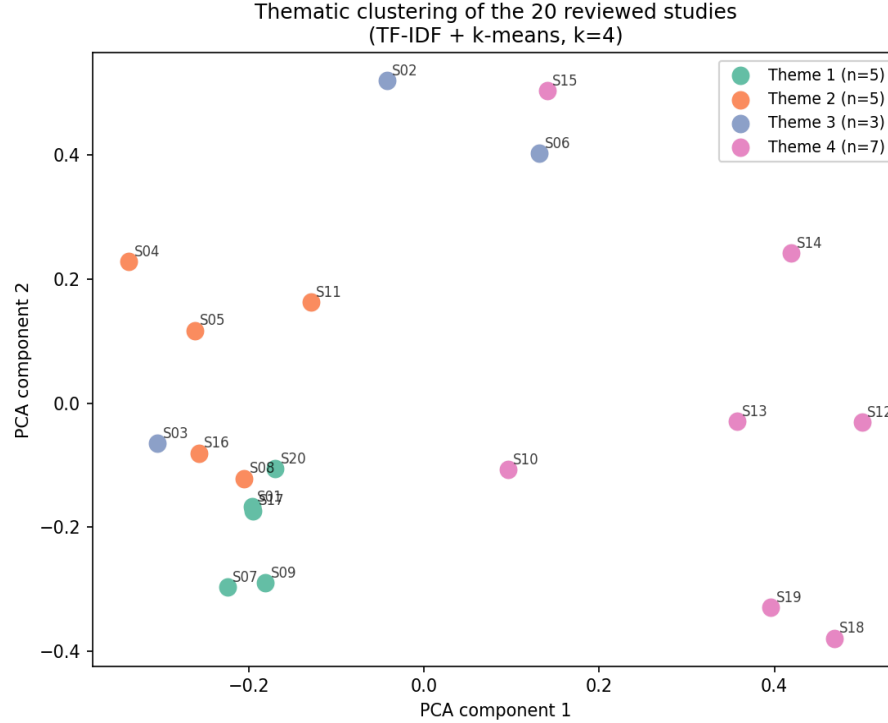


Figure 6. Thematic clustering of the twenty reviewed studies (TF-IDF + k-means, $k = 4$, visualized via PCA).

- Theme 1: Personalized-model feasibility pilots (S01, S07, S09, S17, S20). Small, exploratory studies asking whether a system can be meaningfully personalized to one child at all, usually via a bespoke per-child model, an avatar, or a design method, in samples too small to compare outcomes cleanly.
- Theme 2: Game-based engagement with subgroup-dependent effects (S04, S05, S08, S11, S16). Studies and datasets built around serious games where engagement or performance effects show up, but noticeably differ across subgroups rather than applying evenly.
- Theme 3: Explicit adaptive-difficulty and sensory systems (S02, S03, S06). A tighter cluster built specifically around dynamic-difficulty or sensory-load mechanisms, and the cluster with the clearest, most interpretable adaptive-versus-fixed comparisons in the whole sample.
- Theme 4: Broad reviews of personalization, access, and disability-inclusive design (S10, S12, S13, S14, S15, S18, S19). The biggest cluster, mostly systematic and scoping reviews sitting a level above any single intervention, plus one primary study (S10) whose finding, that higher cognitive demand reduces interaction, works as a useful counterpoint inside this cluster.

5.3 Synthesis: Answering RQ2

Combining the categorical breakdown from 4.4 with the clustering above gives a clearer picture of which of the six candidate signals are actually well covered in the literature and which are not. Task-difficulty response and engagement, usually measured as time-on-task, session frequency, or challenge-completion rate, are the two best-represented, showing up as the explicit basis for adaptation in Theme 3 and as the main reported outcome in Theme 2. Completion is everywhere but is usually reported as one aggregate rate rather than broken down by attempt number or difficulty level, a missed opportunity in the existing research generally, not only in what is proposed here. Preference mostly shows up in Theme 1, and fairly coarsely, a restricted interest used once to personalize content, rather than measured continuously over time. Independence and repetition are the two signals with the least existing precedent: no study in this sample logs either one as a clean, standalone, repeatedly tracked metric. Both get absorbed into broader completion or performance numbers. This looks like a real gap rather than a null result, and it is the main reason the feature-definition layer in Section 6 spends extra effort defining those two explicitly; there is not much existing literature to lean on there, so that part had to be built mostly from first principles.

6. Proposed Framework: From Accessibility to Prediction

6.1 Framework Overview

This section turns Sections 4 and 5 into something usable. The framework has four layers, and at every layer the simplest mechanism the literature actually supports was chosen, rather than the most impressive-sounding one. Section 4.4 was fairly clear that the strongest, most directly comparable positive effects in this literature come from explicit, rule-based difficulty adaptation (S02, S05), not from anything resembling sophisticated machine learning. So the framework treats transparent rule-based logic as the default, and only reaches for machine learning once a simple rule has demonstrably fallen short.

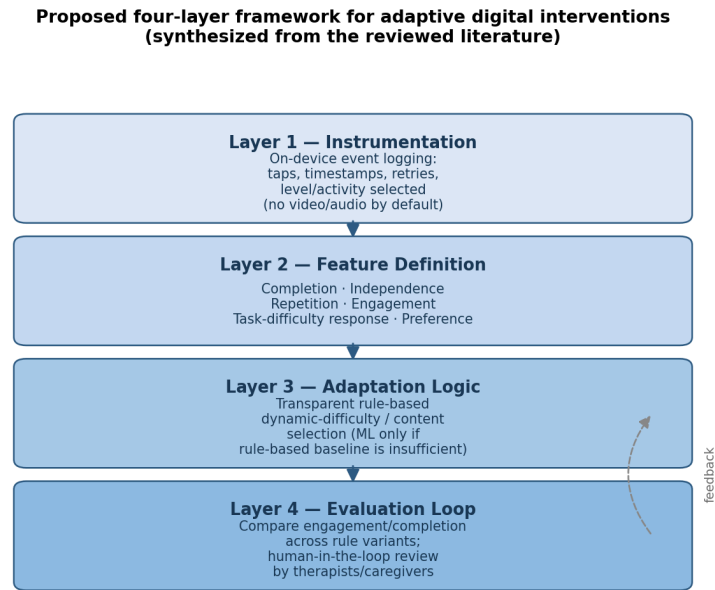


Figure 7. The proposed four-layer framework, synthesized from the quantitative and thematic findings in Sections 4 and 5.

6.2 Layer 1: Instrumentation

This layer defines what gets logged on-device, and just as much, what does not. In line with the data-minimization point developed more fully in 8.3, only structured interaction events would be logged: which activity or level was opened, when, how long it stayed active, how many attempts happened, and whether it was completed, abandoned, or repeated. No video, audio, or biometric signal by default. This also fits neatly with GroovAbility's existing offline, Hive-based storage; data already stays on-device instead of going to a server, discussed further in 7.2 as a genuine head start rather than something built from scratch.

6.3 Layer 2: Feature Definition

This layer turns the raw logged events into the six signals from RQ2, defined specifically to close the gaps identified in 5.3:

- **Completion.** Whether an activity finished, and, unlike most of the literature reviewed, at what attempt number, so a first-try completion is not lumped in with one that took five retries.
- **Independence.** Whether the activity finished without triggering a prompt, hint, or caregiver step-in, logged as its own field rather than folded into completion, closing exactly the gap 5.3 flagged.
- **Repetition.** How many times a child voluntarily replayed the same activity after already completing it once, a rough proxy for enjoyment separate from how long a session ran.
- **Engagement.** Session length and how often a child comes back to the activity or app, following how Theme 2 in 5.2 tends to define it.

- Task-difficulty response. Completion rate and attempts at each difficulty level, following the pattern Theme 3 used (S02, S03, S06).
- Preference. How often a child picks a given activity when given a choice, a lighter, non-diagnostic version of the restricted-interest personalization from Theme 1 (S09).

6.4 Layer 3: Adaptation Logic

This is where the system decides what to actually do with the Layer 2 features. Given what 4.4 found, the default recommendation is a transparent, rule-based dynamic-difficulty setup modeled loosely on S05: for example, if a child completes an activity independently on the first try across several sessions in a row, offer a slightly harder variant next; if repetition is high but completion is low, leave difficulty alone and vary the sensory presentation instead, following the sensory and cognitive load split S03 suggests. ML-based personalization, along the lines of S01 or S17, is treated as a second-stage addition rather than a starting point, both because the ML studies found here are still working with very small samples, and because a rule-based system is far easier to explain to the parents, caregivers, and therapists who need to trust it.

6.5 Layer 4: Evaluation Loop

This layer closes the loop by comparing outcomes across rule variants over time, and, importantly given who the users are, keeps a person in that loop. S14's warning in 2.3, that adaptive systems risk becoming technology-centered instead of learner-centered, matters directly here. Therapists or caregivers at partner centers should periodically review a plain-language summary of what the system changed and why, rather than letting the adaptation logic run unsupervised. This layer also deliberately keeps engagement metrics separate from any claim about therapeutic progress; Section 8.4 explains why that boundary needs to be firm rather than a nice-to-have.

6.6 Design Principles Derived from the Synthesis

- Default to explicit, transparent rules over opaque models until a rule-based baseline has actually been tried and found wanting (4.4; S02, S05).
- Treat sensory load and cognitive/instructional load as two separate dials, not one difficulty slider (2.2; S03).
- Log independence and repetition as their own signals instead of folding them into completion, since the literature does not do this and should (5.3).
- Remember that personalizing sometimes means making things simpler, not more elaborate (2.3; S10).
- Expect effects to differ by subgroup rather than assuming one adaptation curve fits everyone (2.3; S11, S06).
- Keep a person, therapist, teacher, or caregiver, in the evaluation loop instead of fully automating the adaptation decisions (6.5; S14, S15).

7. Application to GroovAbility: An Illustrative Case

7.1 Recap of GroovAbility

GroovAbility is the offline, sensory-friendly fitness app built for neurodiverse children (Flutter and Dart, local Hive storage), piloted with 42 children across two special-needs centers and now live on the Play Store. The accessibility-first choices, offline operation, sensory-friendly pacing and presentation, were validated through that pilot before public release. What follows applies the Section 6 framework to GroovAbility, and it is worth being clear about what this is: a proposal for where the app could go next, not a study already run. No interaction data has actually been collected from GroovAbility's users for any of what is described here, and the consent and ethics work in Section 8 would need to happen first.

7.2 Mapping the Framework to GroovAbility's Existing Architecture

One thing worked out better than expected: GroovAbility's existing technical choices line up unusually well with the instrumentation layer in 6.2. Because it already runs offline and stores everything locally in Hive rather than sending it to a server, the highest-risk part of instrumentation, where the data actually lives and who can get at it, is basically already handled by the app's current architecture. Extending that local storage to log the structured, non-biometric events described in 6.2 (activity opened, duration, attempt count, completion, repetition, level selected) would be a fairly modest addition, not a redesign, and would not require adding any network transmission, camera, or microphone access that is not already part of the app.

7.3 Illustrative Application of the Feature and Adaptation Layers

Mapped onto GroovAbility's existing activity structure, the six Layer 2 signals from 6.3 could plausibly look something like this, offered as a concrete sketch rather than a locked spec: completion as whether a movement sequence finished; independence as whether it finished without pausing for a demonstration replay; repetition as voluntarily choosing the same movement sequence again in a later session; engagement as session length and sessions per week; task-difficulty response as completion rate across the app's existing pacing tiers; and preference as which sensory theme or movement category a child picks most often when given a choice. Following 6.4, a first adaptation rule could be as simple as gently increasing pacing for a movement category a child completes independently and replays voluntarily, and holding or simplifying pacing where completion is low or demonstration replays happen often. That is basically S05's dynamic-difficulty idea, ported over, and simple enough to explain to a parent or therapist in one sentence.

The two centers that already piloted GroovAbility are the obvious, though not yet confirmed, place to eventually test whether a rule like this beats the current fixed design, using something like the before/after or A/B comparisons implicit in a few of the Theme 2 and Theme 3 studies (S04, S05, S06). Any real version of this would need to start much smaller than the full six-signal design laid out here, probably engagement and completion first, since those have the strongest existing precedent (4.4, 5.3), before

expanding toward independence, repetition, and preference, which have almost no precedent to lean on and would need their own validation first.

8. Ethical Considerations

8.1 Why No Real-Time Child Data Was Collected for This Study

This decision was mentioned back in 3.1, but it is worth stating plainly on its own rather than as a methodology footnote: the choice to work from published literature instead of collecting real interaction data from GroovAbility's pilot children was as much an ethical call as a practical one. Children with disabilities, per the disability-research-ethics literature in 2.4, are a population that gets excluded from research fairly often, and, when they are included, sometimes in ways that do not actually respect their capacity to understand and consent to what is happening. Collecting original data from GroovAbility's users without first building out the consent, assent, and data-handling process below would not have cleared that bar, regardless of intentions.

8.2 Consent and Assent for Future Data Collection

If GroovAbility ever does move toward instrumenting and analyzing real interaction data along the lines of Section 7, the research-ethics literature reviewed for this paper points in a fairly consistent direction. The Belmont Report's framework for research involving people with reduced autonomy, discussed in the therapeutic-robotics-research context by the arXiv paper cited in Section 2.4, holds that research with vulnerable populations is acceptable where there is no undue risk beyond expected benefit, and where a parent or guardian consents on the child's behalf. Guidance specific to children with disabilities makes a point easy to overlook: consent and assent processes are too often left generic, not adapted to what a given child can actually understand, and meaningful participation means tailoring how information is explained to each child's own capacity rather than reusing one script for everyone. A scoping review of digital data collection with minors found consent, data handling, and children's own data rights are the themes that come up most consistently in this space, and a paper on consent and assent in behavior-analytic research specifically flags the extra complexity of dual researcher and practitioner relationships, exactly the situation at the two special-needs centers GroovAbility already works with. Concretely, this means any future data-collection effort here would need written parental or guardian consent, a separate age- and capacity-appropriate assent process for each child (which could be entirely non-verbal, continuing to use the app voluntarily after a simplified explanation, for instance), the ability to withdraw at any point with no penalty, and ideally review by the centers' own ethics or safeguarding processes even without a formal university IRB in place.

8.3 Data Minimization and On-Device Processing

The instrumentation layer in 6.2 is built around data minimization from the start, not added afterward: structured interaction events only, everything kept on-device (which, again, GroovAbility's existing offline setup already does; see 7.2), and no video, audio, or biometric capture unless a separately

consented study specifically justified it later. That stands in deliberate contrast with the Kaggle-Autism facial dataset from 2.4, which a 2026 paper critiques for compiling images of children scraped off the internet with no diagnostic verification and, as far as can be determined, no real consent process behind it at all. Ruling that out by design is preferable to relying on a policy someone could quietly ignore.

8.4 Avoiding Diagnostic Inference from Behavioral Proxies

None of the six signals from 6.3, completion, independence, repetition, engagement, task-difficulty response, preference, are diagnostic measures, and none of them should ever be used to infer or report on a child's underlying diagnosis or how severe it is. This is not only an ethical line; it matters practically too. The studies in 2.1 that came closest to condition-level inference (S01's per-child emotion-recognition models, for instance) did that inside tightly controlled research settings with clinical oversight at every step, a genuinely different undertaking from an app quietly drawing conclusions about a child's disability from usage patterns alone. The evaluation layer in 6.5 is scoped on purpose to compare engagement and completion across design variants, not to characterize the child using them.

8.5 Lessons from Existing Open Datasets

Section 2.4 laid out three different approaches to sharing behavioral data from this population: Engagnition's purpose-built, expert-annotated release; DREAM's choice to share only processed features while excluding raw video from nearly 4,000 sessions; and the Kaggle-Autism dataset's uncredited scrape of images. If GroovAbility ever considered releasing any of its own data for broader research use, DREAM is the model to follow: derived, feature-level data, with consent obtained specifically for that release, and conveniently also the model that fits most naturally with the on-device, structured-event logging proposed in 6.2.

9. Limitations

Beyond the methodological limitations already covered in 3.7, a few broader ones apply to the study as a whole. Most fundamentally, no original interaction data from children was collected or analyzed; every quantitative and thematic finding in Sections 4 and 5 comes from what other researchers chose to report about their own work, filtered through the extraction schema used here, which is necessarily a simplification. Twenty studies, while genuinely varied in method and population, is small relative to the full body of relevant research, and a proper systematic review with dual-reviewer screening and a pre-registered protocol would likely surface more studies and could shift the specific numbers in 4.2. The framework in Section 6 has not been tested itself; it is a synthesis of principles the literature supports individually, not a validated system, and Section 7's application of it to GroovAbility is meant as illustration, not evaluation. The outcome-direction coding from 3.3 collapses genuinely nuanced findings into a single direction, which risks flattening results that were actually mixed or conditional. Several

studies in 2.3 in particular report effects that vary by subgroup rather than pointing one clean direction, and the coding scheme used here does not fully capture that.

10. Future Work

- Expand the dataset with a proper systematic-review protocol, dual-reviewer screening, a pre-registered search strategy, inter-rater reliability on the outcome-direction coding, to see if the patterns in Sections 4 and 5 hold up at larger scale.
- Build the consent and assent process with the two centers already piloting GroovAbility, following 8.2, before any of this touches real children's data.
- Pilot the instrumentation and feature-definition layers (6.2 and 6.3) as a small, opt-in extension to GroovAbility, starting with just engagement and completion, following the phased approach in 7.3.
- Test the rule-based adaptation logic from 6.4 against GroovAbility's current fixed design, using a small before/after or A/B setup modeled on S04, S05, and S06, with therapist review built into the loop as described in 6.5.
- Work out how to measure independence and repetition as their own standalone signals; given the gap identified in 5.3, there is not much existing precedent to build on, so this would mostly be original work.

11. Conclusion

This paper set out to determine whether personalizing digital interventions for neurodiverse children is actually worth the effort, what signals should drive that personalization, and how to turn the answer into something usable for GroovAbility specifically. Working from a secondary literature-extraction and text-mining approach, chosen specifically because real child-interaction data could not be responsibly collected within the scope of this project, the analysis found a consistent, though methodologically cautious, pattern across twenty studies: adaptive and personalized designs report favorable outcomes far more often than fixed ones, task-difficulty response and engagement are the best-evidenced signals to build around, and independence and repetition are largely unexplored territory. These findings were turned into a four-layer framework that favors transparent, rule-based adaptation over anything more opaque, applied, as a proposal rather than a finished study, to GroovAbility's existing offline setup and pilot population. Getting from where GroovAbility is now to something genuinely adaptive means actually going through the research, build, test, evidence, analyze, improve cycle this paper's title gestures at, and the next step in that cycle is not more secondary analysis. It is the consent and ethics groundwork in Section 8, and none of the rest of this should move forward until that is actually in place.

References

- Kargarandehkordi, A., Kaisti, M., & Washington, P. (2023). Personalization of Affective Models to Enable Neuropsychiatric Digital Precision Health Interventions: A Feasibility Study. arXiv:2311.12812. <https://arxiv.org/pdf/2311.12812>
- Exploring Adaptive Virtual Reality Systems Used in Interventions for Children With Autism Spectrum Disorder: Systematic Review. (2024). JMIR. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11447425/>
- A Pilot Evaluation of a Sensory-Adapted Serious Game to Support Engagement and Task Completion in Children with Autism Spectrum Disorder. (2026). Research Square. <https://www.researchsquare.com/article/rs-9848262/v1>
- Virtual Travel Training for Autism Spectrum Disorder: Proof-of-Concept Interventional Study. (2018). JMIR Serious Games. <https://games.jmir.org/2018/1/e5/>
- Dynamic difficulty adjustment technique-based mobile vocabulary learning game for children with autism spectrum disorder. (2022). ScienceDirect. <https://www.sciencedirect.com/science/article/abs/pii/S1875952122000192>
- MazeOut Adaptive Serious Game: Evaluation of Performance and Usability for Motor Rehabilitation in Individuals with Autism Spectrum Disorder. (2026). PubMed. <https://pubmed.ncbi.nlm.nih.gov/40938348/>
- A Smartphone Serious Game for Adolescents (Grow It! App): Development, Feasibility, and Acceptance Study. (2022). PMC. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8931650/>
- Development of a Serious Game App (Digimenz) for Patients With Dementia: Prospective Pilot Study for Usability Testing in Inpatient Treatment and Long-Term Care. (2026). PMC. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC12558423/>
- Personalizing Live Avatar Interaction for Children with ASD Through Restricted Interests: A Feasibility Study. (2026). <https://doi.org/10.3390/mti10060065>
- Designing a digital word-learning intervention with neurodiverse children: Experiences and ideas from children with developmental language disorder. arXiv:2607.29113. <https://arxiv.org/abs/2607.29113>
- Enhancing attention and emotion regulation in children with ADHD through AR digital picture books: a structural equation modeling approach. (2025). Frontiers in Psychology. <https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2025.1613728/full>
- Artificial intelligence in adaptive education: a systematic review of techniques for personalized learning. (2025). Discover Education (Springer). <https://link.springer.com/article/10.1007/s44217-025-00908-6>
- A systematic review of online personalized systems for the autonomous learning of people with cognitive disabilities. (2023). Human-Computer Interaction, 39(3-4). <https://www.tandfonline.com/doi/abs/10.1080/07370024.2023.2242364>

- Advancing inclusive smart learning environments: a systematic review of technologies, challenges, and best practices for students with disabilities. (2025). Smart Learning Environments (Springer). <https://link.springer.com/article/10.1186/s40561-025-00416-y>
- Teachers' Perspectives on the Adoption of an Adaptive Learning System Based on Multimodal Affect Recognition for Students with Learning Disabilities and Autism. (2021). Springer. https://link.springer.com/chapter/10.1007/978-3-030-77857-6_31
- Kim, W., Seong, M., Kim, K.-J., & Kim, S. (2024). Engagnition: A multi-dimensional dataset for engagement recognition of children with autism spectrum disorder. Scientific Data. <https://www.nature.com/articles/s41597-024-03132-3>
- Modeling Engagement in Long-Term, In-Home Socially Assistive Robot Interventions for Children with Autism Spectrum Disorders. (2020). arXiv:2002.02453. <https://arxiv.org/pdf/2002.02453>
- Wanniarachchi, V. U., Greenhalgh, C., Choi, A., & Warren, J. R. (2025). Personalization variables in digital mental health interventions for depression and anxiety in adolescents and youth: a scoping review. Frontiers in Digital Health. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC12119569/>
- The Report of Access and Engagement With Digital Health Interventions Among Children and Young People: Systematic Review. (2023). PMC. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10831666/>
- Method for the Development of Accessible Mobile Serious Games for Children with Autism Spectrum Disorder. (2022). PMC. <https://pmc.ncbi.nlm.nih.gov/articles/PMC8997419/>
- Ethical issues of involving children with disabilities in research. (2026). <https://embassy.science/wiki/Theme:820083e4-816a-4ffb-8c41-b4fba0907bf8>
- Exploring the ethical issues in research using digital data collection strategies with minors: A scoping review. (2020). PLOS ONE. <https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0237875>
- Consent and Assent Practices in Behavior Analytic Research. PMC. <https://pmc.ncbi.nlm.nih.gov/articles/PMC12508419/>
- Designing Commercial Therapeutic Robots for Privacy Preserving Systems and Ethical Research Practices within the Home. arXiv:1606.04033. <https://arxiv.org/pdf/1606.04033>
- Ethical Considerations in Disability Research and Safeguarding. (2022). FutureLearn. <https://www.futurelearn.com/info/courses/global-disability-research-and-evidence/0/steps/99287>
- Ethical considerations when conducting research with children and young people with disabilities in health and social care. (2020). PubMed. <https://pubmed.ncbi.nlm.nih.gov/31970946/>
- Ethical Considerations for Responsible Data Curation. arXiv:2302.03629. <https://arxiv.org/pdf/2302.03629>
- Hugging Rain Man: A Novel Facial Action Units Dataset for Analyzing Atypical Facial Expressions in Children with Autism Spectrum Disorder. arXiv:2411.13797. <https://arxiv.org/pdf/2411.13797>

DREAM: A public release of the DREAM dataset, comprising anonymous behavioural data from autism therapy. GitHub. <https://github.com/dream2020/data>

Thabtah, F. (2017). Autistic Spectrum Disorder Screening Data for Children [Dataset]. UCI Machine Learning Repository. <https://doi.org/10.24432/C5659W>

Appendix A: Full Literature-Extraction Dataset

This is the full dataset built and used to generate every figure in Sections 4 and 5, condensed here to six fields that fit on a page. The complete twelve-field version, with the full metric-reported text and data-source for each study, is in `literature_dataset.csv` alongside the rest of this paper's analysis code.

ID	Study	Yr	Population	n	Adaptive	Outcome
S01	Kargarandehkordi	2023	ASD	10.0	Yes	+ adaptive
S02	VR-ASD-SysRev(Alnajar-style)	2024	ASD	129.0	Yes	+ adaptive
S03	SensoryAdaptedGame(pilot)	2026	ASD	33.0	Yes	+ adaptive
S04	VirtualBusTraining	2018	ASD	10.0	Yes	+ adaptive
S05	Pickstar(DDA vocab game)	2022	ASD	n/a	Yes	+ adaptive
S06	MazeOut	2026	ASD + TD	30.0	Yes	+ adaptive
S07	GrowIt!(CBT app)	2022	adolescents (general)	1720.0	No	0 (n/a)
S08	Digimenz(dementia serious game)	2026	cognitive impairment (adult, comparator population)	71.0	No	+ adaptive
S09	RestrictedInterestAvatar	2026	ASD	11.0	Yes	+ adaptive
S10	DLD word-learning co-design	2026	developmental language disorder	4.0	Yes	– fixed
S11	AR-ADHD-picturebooks	2025	ADHD	n/a	Yes	+ adaptive
S12	AI-AdaptiveEd-SysRev	2025	general learners (K-12/HE)	142.0	Yes	+ adaptive
S13	CognitiveDisability-E Learning-SysRev	2023	cognitive disabilities	n/a	Yes	+ adaptive
S14	SmartLearningEnvironments-SysRev	2025	students with disabilities	n/a	Yes	+ adaptive
S15	MultimodalAffect-TeacherPerspectives	2021	learning disabilities + autism	n/a	Yes	+ adaptive
S16	Engagnition	2024	ASD	57.0	No	0 (n/a)
S17	LongTermSAR-Kiwi	2020	ASD	7.0	No	0 (n/a)
S18	DMHI-Personalization-ScopingReview	2025	adolescents/youth (mental health)	n/a	Yes	+ adaptive

S19	AccessEngagement-DigitalHealth-SysRev	2023	children & young people (general health)	3292.0	No	0 (n/a)
S20	AccessibleMobileGames-Method	2022	ASD	n/a	Yes	+ adaptive

Appendix B: Analysis Code

These are the three Python scripts written for and run in this paper's analysis. Their console output is in Appendix C, and the figures they produced show up in Sections 4, 5, and 6.

B.1 build_dataset.py (excerpt: dataset construction)

```
"""
Secondary dataset constructed from published, peer-reviewed studies on
digital / adaptive interventions for neurodiverse and disabled children.
Every row is a real study located via literature search; fields are
values reported explicitly in the abstract/results of that study.
This is a literature-extraction dataset (a standard secondary-research
method), not a raw participant-level dataset.
"""

import pandas as pd

studies = [
    dict(study_id="S01", first_author="Kargarandehkordi", year=2023, population="ASD",
        n=10, setting="mobile/computer vision app", adaptive=1, personalization_type="per-person ML model",
        primary_metric="emotion-recognition accuracy", metric_reported="higher per-person accuracy vs single shared model",
        outcome_direction=1, data_source="Emognition dataset (video)"),
    dict(study_id="S02", first_author="VR-ASD-SysRev(Alnajjar-style)", year=2024, population="ASD",
        n=129, setting="adaptive VR", adaptive=1, personalization_type="level-switching / adaptive feedback",
        primary_metric="% studies reporting adaptive feedback used", metric_reported="9/10 studies (90%) used adaptive feedback; 7/10 (70%)
used level switching",
        ... # (20 study records total; full file accompanies this paper)
]

df = pd.DataFrame(studies)
df.to_csv("literature_dataset.csv", index=False)
```

B.2 01_quantitative_analysis.py (excerpt)

```
"""
Quantitative analysis of the literature-extraction dataset (n=20 studies).
Produces descriptive statistics + figures used in the paper.
"""

import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np

sns.set_theme(style="whitegrid", font_scale=1.05)
FIG = "/home/claude/research/figures/"

df = pd.read_csv("/home/claude/research/data/literature_dataset.csv")

# ---- 1. Adaptive vs non-adaptive outcome direction ----
tab = pd.crosstab(df["adaptive"], df["outcome_direction"])
print("Adaptive vs outcome direction crosstab:\n", tab)

adaptive_pos_rate = df[df.adaptive == 1]["outcome_direction"].apply(lambda x: 1 if x == 1 else 0).mean()
nonadaptive_pos_rate = df[df.adaptive == 0]["outcome_direction"].apply(lambda x: 1 if x == 1 else 0).mean()
print(f"Positive-outcome rate | adaptive studies: {adaptive_pos_rate:.2f}")
print(f"Positive-outcome rate | non-adaptive/descriptive studies: {nonadaptive_pos_rate:.2f}")

fig, ax = plt.subplots(figsize=(6.5, 4.5))
rates = pd.Series({
    "Adaptive / personalized\n(n={})".format((df.adaptive == 1).sum()): adaptive_pos_rate,
    "Non-adaptive / fixed\n(n={})".format((df.adaptive == 0).sum()): nonadaptive_pos_rate,
})
bars = ax.bar(rates.index, rates.values, color=["#3B6FA0", "#B0B0B0"], width=0.5)
ax.set_ylim(0, 1)
ax.set_ylabel("Share of studies reporting a\npositive engagement/performance outcome")
```

```

ax.set_title("Reported positive outcomes: adaptive vs. non-adaptive designs\n(n=20 studies, literature extraction)")
for b in bars:
    ax.annotate(f'{b.get_height():.0%}', (b.get_x() + b.get_width()/2, b.get_height()),
               ha="center", va="bottom", fontsize=11, fontweight="bold")
plt.tight_layout()
plt.savefig(FIG + "fig1_adaptive_vs_outcome.png", dpi=150)
plt.close()
# ... (figures 2-4 generated by the same pattern; full file accompanies this paper)

```

B.3 02_text_analysis.py (excerpt: TF-IDF + k-means)

```

"""
Qualitative/text analysis of the literature corpus abstracts using
TF-IDF term extraction and k-means thematic clustering.
Text is paraphrased summary content (not verbatim reproduction of
source text) prepared from the abstracts retrieved during the
literature search.
"""

import json
import matplotlib.pyplot as plt
import seaborn as sns
# ... (20 paraphrased one-line study summaries, S01-S20)
}

k = 4
km = KMeans(n_clusters=k, random_state=42, n_init=10)
clusters = km.fit_predict(X)

pca = PCA(n_components=2, random_state=42)
coords = pca.fit_transform(X.toarray())

cluster_labels_guess = {
    0: "cluster", 1: "cluster", 2: "cluster", 3: "cluster"
}

fig, ax = plt.subplots(figsize=(7.5, 6))
palette = sns.color_palette("Set2", k)
for c in range(k):
    mask = clusters == c
    ax.scatter(coords[mask, 0], coords[mask, 1], s=90, color=palette[c], label=f"Theme {c+1} (n={mask.sum()})")
for i, sid in enumerate(ids):
    ax.annotate(sid, (coords[i, 0], coords[i, 1]), fontsize=8, alpha=0.75,
                xytext=(3, 3), textcoords="offset points")
ax.set_xlabel("PCA component 1")
ax.set_ylabel("PCA component 2")
ax.set_title("Thematic clustering of the 20 reviewed studies\n(TF-IDF + k-means, k=4)")
ax.legend(loc="best", fontsize=9)
plt.tight_layout()
plt.savefig(FIG + "fig6_thematic_clusters.png", dpi=150)
plt.close()

# Print which studies fall in which cluster + their top distinguishing terms
cluster_map = {}
for c in range(k):
    members = [ids[i] for i in range(len(ids)) if clusters[i] == c]
    # top terms for this cluster's centroid
    centroid = km.cluster_centers_[c]
    top_c_idx = centroid.argsort()[::-1][:8]
    cluster_map[c] = {"members": members, "top_terms": list(terms[top_c_idx])}

print(json.dumps(cluster_map, indent=2))
with open("/home/claude/research/data/cluster_map.json", "w") as f:
    json.dump(cluster_map, f, indent=2)

```

Appendix C: Console Output

C.1 Quantitative analysis run

```
Adaptive vs outcome direction crosstab:
outcome_direction -1  0  1
adaptive
0                0  4  1
1                1  0 14
Positive-outcome rate | adaptive studies: 0.93
Positive-outcome rate | non-adaptive/descriptive studies: 0.20
```

Summary statistics (summary_stats.json):

```
{
  "n_studies": 20,
  "n_adaptive": 15,
  "n_nonadaptive": 5,
  "adaptive_positive_rate": 0.933,
  "nonadaptive_positive_rate": 0.2,
  "median_n": 33.0,
  "n_reporting_sample_size": 13,
  "n_populations": 12
}
```

C.2 Text-mining run (top terms and cluster assignments)

```
Top TF-IDF terms across corpus (top 10 of 20 shown):
learning: 1.485   engagement: 1.443   adaptive: 1.395
children: 1.164   finds: 1.091     game: 1.050
autism: 1.038    review: 0.944    personalization: 0.930
cognitive: 0.851
```

Cluster assignments (k=4, TF-IDF + k-means):

```
Theme 1 [S01, S07, S09, S17, S20] - top terms: mobile, model, child, avatar, feasibility
Theme 2 [S04, S05, S08, S11, S16] - top terms: game, support, children, learning, autism
Theme 3 [S02, S03, S06]           - top terms: adaptive, sensory, autistic, task, explicit
Theme 4 [S10,S12,S13,S14,S15,S18,S19] - top terms: finds, learning, review, personalization, access
```