

Read this first. This document is a format mock-up. The study, the cohort (“Harborlight Middle School”), every number, interval and test statistic below are **simulated** to show what a FLIQ Score™ research paper will look like when a real cohort has been measured. Nothing here is a finding. The measure, its scale, its bands and its eight dimensions are real and are described accurately; the scoring coefficients of the FLIQ Score Model (FSM) are not disclosed here or anywhere.

Measuring change in children’s financial decision-making with the FLIQ Score™: a single-arm pre-post cohort study with 12-month follow-up in one U.S. middle school

An illustrative application of the FLIQ Score Model (FSM) to a Financial Foundations™ cohort, grades 6–8

Adrienne R. Castellanos^{1,*}, Jonah M. Okafor², Lena T. Bergström³, Priya S. Raghunathan², Marcus D. Whitfield⁴, et al.[†]

¹FLIQ Score Research Program, WealthWise Kids LLC, Atlanta, Georgia, United States · ²Department of Educational Measurement, Meridian State University, Georgia, United States (illustrative) · ³Office of Research and Evaluation, Harborlight School District (illustrative) · ⁴Independent evaluator (illustrative)

*Corresponding author: research@wealthwisekids.org · [†]All author names are placeholders. A real cohort paper would list its actual investigators and co-authoring partners under a written data-use agreement.

Article type Original research (illustrative mock)

Study design Single-arm, pre-post cohort with 3-, 6- and 12-month follow-up

Registration Pre-registration template on file; a real cohort would be registered before baseline (OSF or AsPredicted)

Status Not peer reviewed · Not submitted · Simulated data

Version Mock v1.1, 19 September 2026 (figures re-simulated; see §3.2)

Correspondence research@wealthwisekids.org · fliqscore.org/research

ABSTRACT

Background — Financial education for children is evaluated almost entirely with knowledge tests. Meta-analytic evidence shows that knowledge gains translate weakly into behavior, which is what programs are ultimately meant to change. Almost no evaluation observes what a child decides when money is at stake, before and after a program, in the same child. The FLIQ Score™ is a behavioral measure that reads financial decision-making from the choices a child makes inside structured scenarios, scored by the FLIQ Score Model (FSM).

Objective — To describe the change in FLIQ Score™ from baseline to program completion in a middle-school cohort receiving the Financial Foundations™ curriculum, the distribution of that change across dimensions and bands, and its retention over 12 months.

Design, setting and participants — Single-arm pre-post cohort in one suburban, Title I-eligible public middle school (illustrative). All 128 enrolled students in grades 6–8 were eligible; 118 completed a baseline reading and 106 (83% of enrolled; 90% of baseline) a matched completion reading after a 32-week program of 24 planned modules, 22 of them delivered, taught by five trained educators.

Main outcome measures — Primary: change in individual FLIQ Score™ — a composite on a 117–936 scale, reported in five bands across eight dimensions of money behavior — between baseline (T0) and completion (T1). Secondary: change by dimension and grade; band migration; retention at 3, 6 and 12 months (T2–T4); implementation fidelity against seven prespecified indicators.

Results (simulated) — In the matched set ($n = 106$) the mean FLIQ Score™ rose from 528 (SD 121) at T0 to 580 (SD 120) at T1, a mean paired change of +52 FLIQ (95% CI 40 to 64; paired $t_{105} = 8.4$; $d = 0.43$). Change exceeded +20 FLIQ in 70 students (66%), fell within ±20 FLIQ in 24 (23%) and was below –20 FLIQ in 12 (11%). On the 0–100 dimension indices, gains were largest in Borrowing (+12), Saving (+11) and Spending (+10) and smallest in Giving and Protecting (+5 each). The share of students in the Thinker band or above rose from 31 of 106 (29%) to 48 of 106 (45%). At 12 months ($n = 84$) the mean was 567 FLIQ, retaining 76% of that subgroup’s own completion gain. Fidelity composite was 94%.

Conclusions — In this illustrative cohort the FLIQ Score™ recorded a moderate, broadly distributed and largely retained change in measured financial decision-making. Three limits bound that statement. The design has no comparison group, so the change cannot be attributed to the program rather than to maturation or the school year. The T0 and T1 readings share instrument content, so part of the change may be familiarity with the scenarios. And the reliability of the composite has not been estimated, so the confidence intervals reported here assume a precision that has not yet been demonstrated. The purpose of this paper is to fix the reporting format, the analysis plan and the disclosure standard that a real, controlled cohort will follow.

Keywords FLIQ Score; FLIQ Score Model (FSM); financial literacy; financial capability; behavioral assessment; K–12; pre-post cohort; implementation fidelity; measurement

1. Introduction

Financial education reaches tens of millions of school-age children each year, and the evidence for it has matured considerably. The most recent meta-analysis of randomized experiments finds positive effects on both financial knowledge and financial behaviors, with the effect on behavior roughly a quarter the size of the effect on knowledge [1]. Earlier syntheses found that knowledge explains little of the variance in what people actually do with money, and that the effect of instruction decays within months [2]. Large-scale field trials in schools show the same pattern at the level of a national program: knowledge and some self-reported behaviors improve, but the link between the two is loose [3].

The measurement problem sits underneath all of this. Nearly every outcome in the literature is a knowledge test, a self-report, or an administrative record collected years later. What a child *decides* when money is on the table — a trade-off, a deferral, a choice to borrow — is rarely observed directly, and almost never before and after a program in the same child. Assessment science has made this observable elsewhere: stealth assessment infers competencies from choices inside a game or simulation [4]; behavioral economics has decades of task-based measures of risk and time preference [5]; and other fields have built reported scales from observed behavior and normed them to a reference population [6, 7].

The FLIQ Score™ (Financial Literacy Intelligence Quotient) was built on that premise. It reads financial decision-making from the choices a child makes inside structured scenarios and simulations, combines them through the FLIQ Score Model (FSM) into a single score on a 117–936 scale, reports that score in five bands (Explorer, Builder, Thinker, Innovator, Pioneer) and profiles it across eight dimensions of money behavior. The public specification of the measure, its validity status against the AERA/APA/NCME *Standards* [8], and its research agenda are published at fliqscore.org. Its scoring coefficients are held under agreement.

This paper is a format mock-up. It exists to show, in full, how a FLIQ Score™ cohort study will be reported: the study design, the measure, the analysis plan, the tables and figures, the fidelity reporting, the limitations, and the declarations. Every number is simulated. The sections that follow should be read as a template, not as evidence.

2. Methods

2.1 Design and reporting

Single-arm, prospective pre-post cohort with three follow-up readings. Reporting follows the TREND statement for non-randomized evaluations of behavioral interventions [9], adapted to a measurement-focused study. A real cohort would be pre-registered with its primary outcome, analysis plan and fidelity thresholds fixed before baseline data collection.

2.2 Setting and participants

One suburban public middle school (grades 6–8; ages 11–14) eligible for Title I funding (illustrative). All enrolled students in participating classes were eligible; there were no exclusion criteria. Participation in the measure is governed by the school's agreement with WealthWise Kids LLC under FERPA's school-official provision; families received notice and could opt out. Of 128 enrolled students, 118 (92% of enrolled) completed a baseline reading and 111 (94% of those read at baseline; 87% of enrolled) completed at least 80% of the required modules. Twenty of the 118 students read at baseline (17%) reported prior exposure to a financial-education program.

2.3 Intervention

The Financial Foundations™ curriculum: 24 modules over 32 weeks (August 2026 – May 2027), delivered in 45-minute sessions by five school educators trained by WealthWise Kids. Planned instructional time was 18 hours (24 modules × 45 minutes); delivered time averaged 16.5 hours (92%), consistent with 22 of the 24 planned modules delivered. Each module pairs direct instruction with a structured decision scenario; the cohort also completed The \$100 Week™, a five-day browser simulation, at the start and end of the program. The curriculum is copyable and is not the object of study; the measure is.

2.4 The measure: FLIQ Score™ and the FSM

The FLIQ Score™ is derived from observed decisions, not from recall items. Across the program, each structured scenario emits a set of decision signals (for example, the share of a windfall deferred, the acceptance or refusal of a borrowing offer at a stated cost, the response to a simulated setback). Signals are continuous where the underlying decision is continuous and are omitted, never zeroed, when a scenario was not encountered. The FLIQ Score Model (FSM) combines the signals into eight dimension indices (Earning, Spending, Saving, Giving,

Borrowing, Investing, Protecting, Planning), each reported on a 0–100 index, and into one composite score on a 117–936 scale. The structure of the model — which classes of signal enter, how dimensions are weighted in principle, how real-world “node” signals from partner platforms may be admitted under the node protocol — is published; the coefficients are not, and no product computes a score locally. The five bands and their boundaries are fixed and published. A score is reported with its reading date, sitting count and a shrinkage-adjusted estimate when the number of observed decisions is small.

Validity status at the time of writing, stated against the five sources of evidence in the *Standards* [8], is: content — specified; response process — instrumented; internal structure — not yet estimated at scale; relations to other variables — not yet studied; consequences — stated limits on use (never a ranking, never a placement decision). This paper's role is to fix the reporting format that those studies will use.

2.5 Procedures and time points

Readings were taken at baseline (T0, August 2026, before Module 1), completion (T1, May 2027, after the final delivered module), and at 3, 6 and 12 months after completion (T2–T4). Each reading is the score computed from the decision signals available at that date. Baseline and completion readings both include a full pass of The \$100 Week™ so that every student has a comparable set of scenarios at T0 and T1.

2.6 Implementation fidelity

Seven indicators were specified before delivery: modules delivered, educator training, instructional time, baseline completion, follow-up completion, required-module completion, and student session attendance, with platform availability recorded as an eighth operational indicator. Each has a target and a met/partial/not-met status. The fidelity composite is the mean of the seven indicator attainments, each expressed as a percentage and capped at 100%; all seven values are printed in §3.6 so that the composite can be recomputed.

2.7 Statistical analysis

The primary analysis is the paired difference T1 – T0 in students with both readings (matched analysis set). We report the mean difference with a 95% confidence interval from a paired *t*-test, a standardized effect size (Cohen's *d* using the pooled SD of T0 and T1), and the proportion of students whose change exceeded +20 FLIQ, fell within ±20 FLIQ, or fell below –20 FLIQ. Dimension change is reported on the 0–100 index with the same statistics. Band migration is reported as a cross-tabulation of T0 by T1 band. Retention is the mean at T2–T4 in students present at each reading, with the share of the T1 gain retained; a sensitivity analysis carries the last observation forward for students lost to follow-up. Heterogeneity by grade is described, not tested, given the sample size. All intervals are illustrative. Analyses would be run in R (version 4.4) with code deposited alongside the de-identified dataset.

2.8 Ethics and data governance

Student data are collected under the school's authority and held by WealthWise Kids LLC as a school official under FERPA; no personal data are collected from children directly under COPPA outside that authority. Students are identified by a one-way pseudonymous identifier; no names, dates of birth or free-text fields enter the dataset. A real cohort study would be reviewed by the partner university's IRB, and research access would run through a written data-use agreement. Scores are never used to rank students or to make placement decisions.

3. Results (simulated)

3.1 Participant flow

Of 128 enrolled students, 118 completed a baseline reading, 111 completed the program and 106 completed a matched T1 reading (Table 1). The twelve students without a T1 reading had transferred schools (4), were absent at follow-up (5) or withdrew (3). Students lost to follow-up scored higher at baseline (mean 534, SD 93) than those retained (mean 528, SD 121), a difference of 6 FLIQ. A real cohort should report this comparison with its direction and magnitude rather than assert equivalence, whichever way it falls.

Table 1. Participant flow and readings by time point (simulated).

Time point	When	n read	% of enrolled	Mean FLIQ	SD
Enrolled	Aug 2026	128	100	—	—
T0 Baseline	Aug 2026	118	92	529	118
T1 Completion	May 2027	106	83	580	120
T2 3 months	Aug 2027	98	77	576	122
T3 6 months	Nov 2027	91	71	570	117
T4 12 months	May 2028	84	66	567	117

3.2 Primary outcome

In the matched set ($n = 106$) the mean FLIQ Score™ rose from 528 to 580, a paired change of +52 FLIQ (95% CI 40 to 64; paired $t_{105} = 8.4$, $p < .001$). Using the pooled SD of the matched set (121, from T0 SD 121 and T1 SD 120), $d = 0.43$; the SD of the within-student difference was 64. The median rose from 528 to 580. Change exceeded +20 FLIQ in 70 students (66%), fell within ± 20 FLIQ in 24 (23%) and was below -20 FLIQ in 12 (11%). The baseline mean of the matched set (528) differs from the baseline mean of all students read at T0 (529, $n = 118$, Table 1); the paired analysis uses the matched set throughout. Change by grade was +43 in grade 6 ($n = 36$), +56 in grade 7 ($n = 36$) and +56 in grade 8 ($n = 34$), these are rounded subgroup means and size-weight to +51.6 against an overall +52, the difference being an artefact of rounding each subgroup to the nearest FLIQ; grade is described, not tested.

We note for the reader how this simulated effect was chosen. The meta-analytic average effect of financial education on behavior is near $d = 0.09$, and on knowledge near $d = 0.33$ [1]. Those are controlled contrasts against a comparison group. A single-arm pre-post change of the kind reported here also carries maturation, the school year and familiarity with the scenarios, so it is legitimately larger than a controlled effect; $d = 0.43$ was chosen to sit in that range rather than above it. A template that enshrined an implausibly large effect would set an expectation that no real cohort could meet, and would invite exactly the scrutiny this reporting standard exists to survive.

3.3 Dimensions

All eight dimension indices rose (Table 2). The largest gains were in Borrowing (+12), Saving (+11) and Spending (+10); the smallest in Giving and Protecting (+5 each). Investing remained the lowest index at completion (43), with Protecting next (46). Gain is only moderately related to how much room a dimension had: across the eight indices the correlation between baseline and change is $r = -0.52$ ($r^2 = 0.27$), so headroom accounts for some of the pattern but leaves most of it unexplained. Every interval excludes zero, but these are eight related outcomes tested on one cohort with no multiplicity adjustment; they are descriptive.

Table 2. Dimension indices (0–100) at baseline and completion, matched set, $n = 106$ (simulated).

Dimension	T0	T1	Δ	95% CI
Earning	53	59	+6	4 to 8
Spending	49	59	+10	8 to 13
Saving	44	55	+11	9 to 13
Giving	58	63	+5	3 to 8
Borrowing	38	50	+12	10 to 14
Investing	34	43	+9	6 to 11
Protecting	41	46	+5	3 to 7
Planning	48	56	+8	6 to 11

3.4 Band migration

At baseline 75 of 106 students (71%) were in the Explorer or Builder bands; at completion 48 of 106 (45%) were in the Thinker band or above (Table 3). 34 students moved up at least one band and 4 moved down. Counts and percentages are printed separately throughout: the two are easily transposed, and a band table is where that error hides.

Table 3. Band at baseline (rows) by band at completion (columns), n = 106 (simulated).

T0 \ T1	Explorer	Builder	Thinker	Innovator	Pioneer	Total
Explorer	8	8	1	0	0	17
Builder	2	40	16	0	0	58
Thinker	0	0	18	7	2	27
Innovator	0	0	1	2	0	3
Pioneer	0	0	1	0	0	1
Total	10	48	37	9	2	106

3.5 Retention

Mean scores at 3, 6 and 12 months were 576 (n = 98), 570 (n = 91) and 567 (n = 84). Follow-up samples are nested, each drawn from the one before it. Retention is computed within each subgroup against that subgroup's own baseline and completion means, so that attrition does not masquerade as decay; those means are 527 and 582 at T2, 528 and 584 at T3, and 525 and 580 at T4, so each figure can be recomputed. On that basis 89%, 75% and 76% of the gain was retained. The prespecified sensitivity analysis, carrying the last observation forward for the 106 – 84 students not read at 12 months, gives a matched-set mean of 570 and 81% retention. On the levels, the mean falls at every reading (580 → 576 → 570 → 567), with 10 of the 13-point total fall arriving by six months; whether that trajectory differs from the decay reported for knowledge outcomes [2] is one of the questions a real cohort exists to test. With no comparison group, none of these figures separates retention from ordinary development over the same period.

3.6 Implementation fidelity

Five of seven indicators met their target; two fell short of a full-delivery target. The seven attainments were: modules delivered 22 of 24 planned (target 24), 92%; educator training 5 of 5 educators (target 5), 100%; instructional time 16.5 of 18 planned hours (target 18), 92%; baseline completion 118 of 128 enrolled (target 85%), 92%; follow-up completion 106 of 118 read at baseline (target 85%), 90%; required-module completion 111 of 118 (target 85%), 94%; and student session attendance, the mean share of delivered sessions attended per student (target 90%), 96%. Platform availability was 99.6% of scheduled session hours across the 32-week delivery period, recorded as an eighth operational indicator and excluded from the composite. The fidelity composite – the mean of the seven percentages above – was 94%.

4. Discussion

4.1 Principal findings

In this illustrative cohort the FLIQ Score™ recorded a moderate mean gain that was broadly distributed across students, accompanied by an upward shift across bands and largely retained at 12 months. It was not uniform: 12 students (11%) fell by more than 20 FLIQ, and the dimensions that gained most were not simply those with the most room to move. A reporting format that hid either fact would be the wrong format. The structure demonstrated here – a primary paired outcome with its interval and effect size, dimension-level change, band migration reported in both counts and percentages, retention computed within nested subgroups with a sensitivity analysis, and fidelity against prespecified thresholds with every indicator printed – is the structure every FLIQ Score™ cohort report will use.

4.2 Relation to the literature

The financial-education literature has moved from asking whether programs work to asking what they change and for how long [1, 3]. Its persistent weakness is that behavior is measured indirectly. A behavior-derived score, if it can be shown to be reliable and to relate to later real-world outcomes, addresses that weakness at the level of the instrument. The evidence ladder for such a claim is well established in neighbouring fields: internal consistency and test-retest reliability, then convergent and discriminant relations to existing measures (for example the OECD/PISA financial-literacy assessment [10] and the CFPB financial well-being scale [6]), then predictive relations to observed behavior, then consequences of use. The FLIQ Score™ is at the first rung.

4.3 Strengths and limitations

The strengths of the design are procedural: a prespecified primary outcome, a matched analysis set, fidelity reported against thresholds fixed in advance, follow-up to 12 months, and a stated

disclosure standard. The limitations are fundamental. There is no comparison group, so the change cannot be attributed to the program rather than to maturation, practice with the scenarios, or the school year itself. Readings at T0 and T1 share instrument content, so some gain may reflect familiarity. Reliability of the composite and the dimension indices has not been estimated, so the intervals reported here assume a precision that has not been demonstrated. Twelve of 118 students had no completion reading; their baseline mean differed from the retained group by 6 FLIQ, about 0.05 SD, which is small, but the matched set is still not a random subset and a larger or more selective loss would matter more. The eight dimension comparisons carry no multiplicity adjustment. The band cut points that generate Table 3 are fixed and published but are not printed in this paper, so a reader cannot reproduce that table from the scores alone; a real cohort report should print them. The cohort is one school. And, above all, the data are simulated.

4.4 Implications and next studies

Three studies follow directly. First, a reliability study: internal structure and two-week test-retest of the composite and eight indices in a cohort of at least 300 students, with McDonald's $\omega \geq .70$ as the pre-registered gate. Second, a convergent-validity study relating the FLIQ Score™ to an established knowledge measure and to a task-based measure of time preference, to establish that the score is not simply a proxy for either. Third, a matched-comparison or waitlist-controlled cohort with the design shown here, so that change can be attributed. The FLIQ Score Research Program invites university partners to own any of these under a written data-use agreement, with results published whatever they show.

5. Conclusion

This paper fixes the reporting standard for FLIQ Score™ cohort studies before any real cohort is reported. When the first measured cohort is published, it will use these sections, these tables, these declarations and this disclosure standard, and its numbers will replace the simulated ones here.

Declarations

Competing interests

A.R.C. is employed by WealthWise Kids LLC, which operates the FLIQ Score™ under license from WealthWise Kids IP Holdings LLC and developed the FLIQ Score Model (FSM). M.D.W. is an independent evaluator with no financial interest in the

program and is responsible for the analysis. The remaining authors declare no competing interests. (Illustrative.)

Funding

None for this mock-up. A real cohort would state its funder, the funder's role, and any relationship between the funder and the program.

Ethics approval

Not applicable to simulated data. A real cohort would be reviewed by the partner university's IRB and conducted under the school's FERPA authority.

Data availability

The simulated dataset and the analysis template are available on request from research@wealthwisekids.org. Real cohort data would be shared as a de-identified dataset under the study's data-use agreement.

Author contributions (CRediT)

Conceptualization and methodology: A.R.C., J.M.O. Investigation and data curation: L.T.B., P.S.R. Formal analysis: M.D.W. Writing — original draft: A.R.C. Writing — review and editing: all authors. (Illustrative.)

References

1. Kaiser T, Lusardi A, Menkhoff L, Urban C. Financial education affects financial knowledge and downstream behaviors. *Journal of Financial Economics*. 2022;145(2):255–272.
2. Fernandes D, Lynch JG, Netemeyer RG. Financial literacy, financial education, and downstream financial behaviors. *Management Science*. 2014;60(8):1861–1883.
3. Bruhn M, de Souza Leão L, Legovini A, Marchetti R, Zia B. The impact of high school financial education: evidence from a large-scale evaluation in Brazil. *American Economic Journal: Applied Economics*. 2016;8(4):256–295.
4. Shute VJ. Stealth assessment in computer-based games to support learning. In: Tobias S, Fletcher JD, eds. *Computer Games and Instruction*. Information Age Publishers; 2011:503–524.
5. Lejuez CW, Read JP, Kahler CW, et al. Evaluation of a behavioral measure of risk taking: the Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: Applied*. 2002;8(2):75–84.
6. Consumer Financial Protection Bureau. *Measuring Financial Well-Being: A Guide to Using the CFPB Financial Well-Being Scale*. CFPB; 2015.
7. Cella D, Riley W, Stone A, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of Clinical Epidemiology*. 2010;63(11):1179–1194.
8. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. AERA; 2014.
9. Des Jarlais DC, Lyles C, Crepaz N; TREND Group. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. *American Journal of Public Health*. 2004;94(3):361–366.
10. OECD. *PISA 2018 Results (Volume IV): Are Students Smart about Money?* OECD Publishing; 2020.