Psychometrika (2026), **91**, 424–444
doi:10.1017/psy.2026.10087

PSYCHOMETRIC SOCIETY

THEORY AND METHODS

Causal Mediation and Functional Outcome Analysis with Process Data

Youmi Suk[1] and Chan Park[2]

[1]Teachers College, Columbia University, USA; [2]University of Illinois Urbana-Champaign, USA

Corresponding author: Youmi Suk; Email: ysuk@tc.columbia.edu

(Received 29 January 2025; revised 23 December 2025; accepted 23 December 2025)

Youmi Suk and Chan Park contributed equally to this manuscript as co-first authors. The order of their names was randomly selected.

Abstract

Over the past two decades, there has been growing interest in analyzing the effects of educational programs on outcomes using process data from computer-based testing and learning environments. However, most analyses focus on final outcomes at the end of a test or session, overlooking their functional nature over time and neglecting causal mechanisms. To address this gap, this article proposes a novel causal mediation framework for identifying and estimating functional natural direct effects, functional natural indirect effects, and functional total effects, along with their subgroup effects. We define these effects using potential outcomes and provide nonparametric identification strategies depending on whether post-treatment covariates are present or not. We then develop estimation methods using generalized additive models, a flexible and robust tool for analyzing functional data. Through a simulation study, we assess the finite-sample performance of the proposed approach by comparing it to parametric regression methods. We also demonstrate our approach by examining the effects of extended time accommodations on two functional outcomes using process data from the National Assessment of Educational Progress. Our mediation approach with functional outcomes effectively captures dynamic causal mechanisms underlying the program's effects and pinpoints when and for whom each effect manifests throughout the testing period.

Keywords: causal inference; causal mediation; functional data; generalized additive models; process data

1. Introduction

The transition to computer-based testing and digital assessments in various testing programs began in the early 2000s. Recent technological advancements and evolving educational needs have driven the widespread use of these digital assessments (Bennett et al., 2007; Thurlow et al., 2010). A key feature of digital assessments is the collection of *process data*—information on all of the responses about test-takers' interactions with computer devices during the test. Process data include real-time responses and actions, such as mouse clicks, keystrokes, response times, and behaviors related to testing accommodations (Bergner & von Davier, 2018). Not surprisingly, the rich and detailed information from these data helps researchers understand test-takers' problem-solving behaviors and their engagement with accessibility tools and testing accommodations. While research using process data has been actively conducted for tasks, such as ability estimation and cheating detection (Jiao et al., 2021), its methods and applications to causal inquiry are still in their early stages. The overarching goal of this article is to propose a causal framework with time-stamped process data to assess the dynamic causal effects of test-related or educational programs.

© The Author(s), 2026. Published by Cambridge University Press on behalf of Psychometric Society.
This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/psy.2026.10087 Published online by Cambridge University Press

As a concrete example, consider evaluating the effect of extended time accommodations (ETAs), the most frequently provided accommodation to students with disabilities and/or English language learners. Previous studies (Lee & Suk, 2025; Suk & Kim, 2025; Wei & Zhang, 2023) used process data to identify whether students who received ETA made use of it during a test and, with this variable, evaluated the effects of using ETA on test scores through different quasi-experimental designs (see Section 2.1 for details). However, these studies focus on test-takers' final outcomes measured at the end of a test or session and overlook their functional nature over time. It is possible that ETA treatment groups may not differ in terms of final outcomes, but may show temporal or consistent differences during specific periods of testing. Capturing such changes requires researchers to analyze evolving test scores over time rather than the final test scores and view the former as *functional outcomes* from the field of functional data analysis (Ramsay & Silverman, 2005; Wang et al., 2016). Briefly, functional data are defined as data that change over a continuous domain (e.g., time and space), and each unit's outcome is typically a function or a curve that varies over a continuum rather than a single discrete point. Regression models for such functional outcomes try to capture evolving patterns like time-dependent processes or two-dimensional images (Ramsay & Silverman, 2005) (see Section 2.2 for more details on functional data). However, there is still an open question as to how to analyze process data in a functional format to draw causal inferences.

Moreover, the rich information from process data has the potential to uncover the *causal mechanisms* in digital assessments, i.e., how and why educational programs influence an outcome. For example, previous studies on the ETA evaluation (e.g., Lee & Suk, 2025; Suk & Kim, 2025) assume that the effect of ETA on an outcome manifests only by using it. However, changes in outcomes may result either from using ETA (i.e., indirect effect) or from receiving ETA itself, not through its actual use (i.e., direct effect). This direct effect captures some psychological benefits of receiving ETA, even if the ETA itself is not actually used. For example, the ETA availability can positively affect student performance by reducing anxiety (Lovett & Leja, 2013). The direct and indirect effects may also vary depending on subgroups defined by test-takers' characteristics (e.g., disability status and gender). Investigating both direct and indirect effects and conducting their subgroup/moderation analysis helps gain a more nuanced understanding of program effects. While existing causal tools on causal mediation (e.g., Pearl, 2001) and subgroup analysis (e.g., Yang et al., 2021) can support such investigations, a systematic methodology is still needed to extend these tools to the functional data context, in order to capture dynamic aspects of direct and indirect effects using process data in a functional format.

This article proposes a novel causal mediation framework with a functional outcome to evaluate test-related or educational programs in process data. The proposed framework provides the counterfactual definitions for the functional natural direct and indirect effects, and their total average effect. For subgroup analysis, we define these functional effects among subgroups determined by a subset of pre-treatment covariates. We also provide nonparametric identification strategies for each effect depending on the presence or absence of post-treatment confounders and construct estimation methods using generalized additive models (GAMs; Hastie & Tibshirani, 1986), a flexible and robust tool for analyzing functional data. Through a simulation study, we demonstrate the effectiveness of the proposed GAM-based methods by comparing them with parametric regression methods. Furthermore, we apply our framework to examine the effects of ETA on two functional outcomes—evolving test scores and evolving item access—using process data from the National Assessment of Educational Progress (NAEP).

The remainder of the article is organized as follows. Section 2 reviews the relevant literature and highlights our contributions. Section 3 describes the study setup. Section 4 outlines our causal mediation framework with a functional outcome by building on the existing causal mediation literature. Section 5 describes the design and results of our simulation study, and Section 6 demonstrates our framework to evaluate the effects of ETA in the NAEP process data. Finally, Section 7 discusses the implications and conclusions of this study.

https://doi.org/10.1017/psy.2026.10087 Published online by Cambridge University Press

2. Prior work and our goals

2.1. Causal inference with process data

Most of the research efforts using process data from digital assessments have focused on estimating ability parameters, facilitating cognitive diagnosis, and detecting aberrant responding behavior (Jiao et al., 2021). However, only a few studies have used process data to explore causal questions about students' testing behaviors or test-related programs. For example, Wei and Zhang (2023) use a propensity score method to balance covariates across three distinct profiles of students with learning disabilities: those who do not receive ETA, those who receive ETA but do not use it, and those who both receive and use ETA. They construct students' ETA use status based on the NAEP process data by comparing their total response time with the standard testing time, and then analyze the math performance among the three profiles. However, Suk and Kim (2025) suggest using a fuzzy regression discontinuity design with multiple control groups to evaluate the effects of using ETA near the cutoff of a running variable. In their approach, the ETA eligibility status serves as an instrumental variable, and the effect bounds are constructed using multiple control groups. More specifically, the running variable is the English proficiency levels of English language learners, and multiple control groups occur because students either do not receive ETA or decline to use it when offered. Using the NAEP process data, they construct the ETA use status similarly to Wei and Zhang (2023) and find no strong evidence of using ETA on the final test scores among a subpopulation.

Additionally, Lee and Suk (2025) propose an evidence factor analysis in fuzzy regression discontinuity designs to reinforce causal conclusions on the effect of using ETA within a single dataset. Evidence factors represent two or more independent tests of the same null hypothesis about the treatment effect within a single dataset, each potentially subject to different biases (Rosenbaum, 2010). They also use the NAEP process data to demonstrate how to construct and combine three evidence factors in the ETA example, and as a result, there is no strong evidence of the effect of using ETA on the final test scores.

Although these prior studies have utilized process data to find causal evidence on students' behaviors or testing programs, they have only incorporated a limited portion of process data into causal research, without fully using time information. These studies also assume that the treatment effect manifests only by using it and ignore the potential direct effect of receiving it, not mediated by its actual use. Therefore, there is a need to develop a rigorous methodology for analyzing time-stamped process data to draw causal inferences and discover the underlying mechanisms.

2.2. Causal inference with functional data

In classical statistical methods, the fundamental observational unit is typically a scalar or a vector of fixed length. However, with advances in technologies, observational units can take more complex forms, such as functions, curves, images, or other objects that vary over a continuum. Functional data analysis is a branch of statistics specifically developed to handle such non-scalar data. For simplicity, we will refer to these complex objects as "functions" throughout the manuscript. For example, in NAEP process data, one might traditionally analyze only the final math test score using classical statistical methods, say using a linear regression that regresses the final test score on other covariates. However, if process data capture the *entire trajectory* of the evolving test score over time, it can be viewed as a function over the testing period. Such functional observations can preserve the full temporal dynamics of test-taking processes.

With the shift in the observational unit, the goals of functional data analysis include understanding relationships, making predictions, and identifying patterns with functional variables. In this field, methodological advances have been made to more effectively address the high dimensionality and inherent complexity of functional data. Key developments include functional principal component analysis, functional regression with functional outcomes and/or functional predictors, and clustering of functional data (Crainiceanu et al., 2024; Ramsay & Silverman, 2005; Wang et al., 2016). However,

https://doi.org/10.1017/psy.2026.10087 Published online by Cambridge University Press

most existing studies are limited to exploring associations rather than causal relationships, and the development of causal inference techniques with such functional data is in its infancy.

Recently, some notable work has been developed. Ecker et al. (2023) introduce a causal framework with a functional outcome (e.g., cumulative lifetime incomes) and a scalar binary treatment (e.g., residing in urban versus rural areas) to target the functional average treatment effect (FATE), i.e., the average treatment effect (ATE) over the functional domain. They extend the standard causal assumptions (consistency, positivity, and ignorability/unconfoundedness) to the functional data context to identify the FATE. Ecker et al. (2023) use functional linear regression as the outcome model to estimate the FATE and develop simultaneous confidence bands for valid inference across the functional domain. They also examine the subgroup effects of FATE to find effect heterogeneity, say between males versus females. In contrast, Gao et al. (2024) propose a causal framework with a functional treatment (e.g., hippocampal changes over time) and time-to-event (survival) outcomes (e.g., time until conversion from mild cognitive impairment to Alzheimer's disease). They define the target estimand as the coefficient of functional treatment in a type of functional linear regression model and provide three estimation methods, including regression methods, weighting methods, and combinations of these.

Since process data are collected over a specific time period, it can be considered a type of functional data. As such, causal inference methods for functional data could be applied to evaluate the effects of students' testing behaviors or test-related programs on the functional outcomes, such as evolving test scores. However, we need additional considerations to accurately establish a functional causal framework for process data. For example, process data from digital-based assessments may require additional assumptions for causal identification due to the unique nature of assignment and outcome processes in digital environments (e.g., one-sided noncompliance and constant potential outcomes over a certain period). Furthermore, while existing studies (Ecker et al., 2023; Gao et al., 2024) focus on estimating the FATE (or a type of ATE), they are limited in investigating causal mechanisms through mediators across the functional domain. In cases where researchers wish to decompose the total effect into direct and indirect effects in functional data, they need to develop a new framework and methods for causal mediation with functional data.

2.3. The goals of this article

To address the limitations of existing approaches, this article seeks to develop a new causal mediation framework for functional outcomes derived from process data. Specifically, we pursue the following four goals.

First, we aim to re-conceptualize process data as functional outcomes. While researchers in education and psychology often interpret evolving outcomes (e.g., test scores recorded during a test) as repeated discrete outcomes within a longitudinal framework, we adopt a functional data perspective. Under this paradigm, evolving test scores are treated as realizations of an underlying continuous process, where the study unit is an individual's entire outcome trajectory rather than time-point measurements. This functional view is more appropriate for process data and marks a conceptual shift that is relatively new to the fields of psychology and education. We illustrate this shift using a motivating example of ETA, which serves as the empirical context throughout the manuscript.

Second, we aim to define causal estimands that uncover dynamic causal mechanisms over a functional domain (e.g., time or space). Specifically, we define the average functional indirect and direct effects, along with their subgroups effects, where the subgroups are defined by a subset of pre-treatment covariates (see Equations (1)–(4) for details). The resulting framework enables researchers to effectively answer the causal question: "*Why, when, and for whom does a program work?*" The "*why*" component concerns causal mechanisms through direct and indirect pathways; the "*when*" component captures how effects evolve over the functional domain; and the "*whom*" component addresses subgroup effects beyond the ATE. When the functional domain corresponds to space, our framework can also address *where* a program is effective.

https://doi.org/10.1017/psy.2026.10087 Published online by Cambridge University Press

Third, we aim to establish identification strategies under several realistic scenarios, including: (i) the presence or absence of post-treatment confounders, (ii) mediators derived from other variables, and (iii) one-sided versus two-sided noncompliance. In particular, the presence of post-treatment confounders requires jointly modeling the mediator and post-treatment confounders. We outline the identification assumptions that are most relevant to our process data setting in the main text, while providing detailed identification results for each scenario in the appendices of the Supplementary Material to broaden the applicability of our approach.

Fourth, we aim to develop estimation and inference procedures for functional causal effects using GAMs. GAMs are widely used in functional data analysis due to their flexibility and adaptivity (Crainiceanu et al., 2024; Ramsay & Silverman, 2005). We leverage GAMs to construct estimators for our functional direct and indirect effects. Importantly, our estimators go beyond fitting outcome regression by incorporating carefully designed weighting schemes to account for post-treatment variables. For inference, we employ a cluster bootstrap that resamples individuals, rather than time points, to consider unit-specific serial dependence. Through simulation studies, we demonstrate that our GAM-based methods achieve a favorable bias–variance trade-off and perform at least as well as correctly specified parametric methods.

3. Setting

3.1. Notation

Let $i = 1, \ldots, n$ index subjects, i.e., study units. We suppress the subscript i unless necessary. We consider a setting where the following variables are available for subject i from functional data. We denote the treatment receipt variable as $Z \in \{0,1\}$, where $Z = 1$ indicates that a subject received the treatment and $Z = 0$ indicates that the subject did not receive it. We denote a set of pre-treatment covariates as $X = \{X_1, \ldots, X_p\} \in \mathbb{R}^p$. We also denote the observed outcome process as $Y(\mathcal{K}) = \{Y(k) : k \in \mathcal{K} = [k_0, k_m]\}$, where k_0 and k_m represent the minimum and maximum values of the functional domain $\mathcal{K}$, respectively. The minimum value k_0 plays a key role in the definition and identification of the functional causal effects (see Section 4 for details). Additionally, $W = \{W_1, \ldots, W_q\} \in \mathbb{R}^q$ denotes the observed post-treatment covariates by time k_0 (e.g., the number of accessed items up to k_0).

For causal mediation analysis with functional data from digital assessments, we introduce additional variables. Let $T \in \{0,1\}$ denote the actual treatment use, where $T = 1$ indicates that a subject used the treatment and $T = 0$ indicates that the subject did not use it. Here, T and Z differ because subjects who receive the treatment can decide whether to use it, and T serves as a mediator. We denote a subject's total response time as $H \in \mathcal{K}$. In particular, we observe the outcome process $Y(k)$ over $k \in [k_0, H]$, where the last observation ends at time H, and we use H to derive the treatment use variable as $T = \mathbb{1}\{H > c\}$. Therefore, the observed data for a subject consists of $(X, Z, T, H, W, Y([k_0, H]))$ (see Table 1 for the structure of the process data). In Table 1, the first subject ($i = 1$) has three covariates $\{X_1, X_2, X_3\}$ with values $\{1, 0, 0\}$, an ETA receipt status of $Z = 0$ (did not receive), an ETA use status of $T = 0$ (did not use), and a total response time of $H = 5$ minutes. Under $k_0 = 3$, this student's post-treatment covariate is observed as $W = 2$. Their functional outcome Y is observed from $k_0 = 3$ to their total response time H, with values of 1, 2, and 3 at $k = 3,4$, and 5 minutes, respectively.

Following the Neyman–Rubin's potential outcome framework (Neyman, 1923; Rubin, 1974) and its extension to causal mediation (Pearl, 2001; Robins & Greenland, 1992), we define potential (also known as counterfactual) variables. Let $T^z(k)$ denote the potential treatment use (i.e., mediator) if Z were set to $Z = z$ at time k. The relationship between treatment Z and mediator T can be either two-sided or one-sided. For two-sided noncompliance, we assume that $T^z(k) \in \{0,1\}, z = 0,1$, but for one-sided noncompliance, we assume that $T^1(k) \in \{0,1\}$ and $T^0(k) = 0$. Similarly, let W^z be a subject's potential post-treatment covariate if they received the treatment $Z = z$. This implicitly assumes that the post-treatment covariate W is not affected by the treatment use T, but W influences T (see Figure 1 in the

https://doi.org/10.1017/psy.2026.10087 Published online by Cambridge University Press

Table 1. An example of the process data with three-dimensional $X = \{X_1, X_2, X_3\}$, one-dimensional W, and $k_0 = 3$

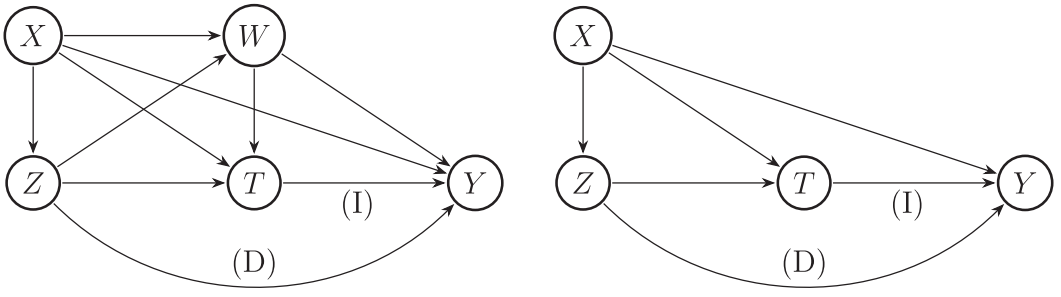
i	$X = \{X_1, X_2, X_3\}$	Z	T	H	W	$Y(\{3,4,\ldots,H\})$
1	{1,0,0}	0	0	5	2	{1,2,3}
2	{0,1,1}	1	0	4	1	{1,2}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	{1,1,1}	1	1	7	1	{0,1,2,2,3}

next section for a graphical illustration). Lastly, let $Y^{z,T^{z'}}(\mathcal{K}) = \{Y^{z,T^{z'}(k)}(k) : k \in \mathcal{K}\}$ denote a subject's potential outcome process if Z were set to z and T were set to the value it would naturally take under $Z = z'$. For example, $Y^{1,T^0}(\mathcal{K})$ indicates the potential outcome process if a subject received the treatment (i.e., $Z = 1$) and their mediator T took the value it would have attained under the control condition (i.e., $Z = 0$).

3.2. Motivating example

Testing accommodations are essential for students with disabilities and English language learners to fully demonstrate their abilities during a test. The ETA is the most frequently provided testing accommodation in many testing programs. Suppose that a researcher is interested in studying the causal mechanism and determining whether the benefit of ETA is a consequence of using ETA (i.e., indirect effect) or partially attributed to receiving ETA, not mediated via its actual use (i.e., direct effect). We also observe one-sided noncompliance in our ETA example; only students who receive ETA can decide to use it or not, but those who do not receive ETA cannot use it.

Figure 1 visualizes the assumed causal mechanisms of ETA with the receipt status (Z) and use status (T). Figure 1a describes the case where post-treatment covariates W are present, and Figure 1b describes the case only with pre-treatment covariates X. Confounding between T and Y arises due to students' decisions about whether to use ETA, influenced by Z, X, and, if applicable, W. The total effect of Z on functional outcome Y can be decomposed into the indirect causal pathway mediated by T (i.e., $Z \to T \to Y$) and the direct pathway not mediated by T (i.e., $Z \to Y$).



(a) Post-treatment covariate W (b) Pre-treatment covariate X only

Figure 1. Graphical representations of causal relationships with a functional outcome. The arrows labeled with (D) and (I) represent the functional direct effect and functional indirect effect, respectively.

https://doi.org/10.1017/psy.2026.10087 Published online by Cambridge University Press

4. Causal mediation with a functional outcome from process data

In this section, we extend the standard causal mediation framework (e.g., Pearl, 2001) to the functional outcome context, particularly using process data in digital-based assessments. We use the ETA in digital-based assessments as an illustration of the proposed framework. In this setting, we observe one-sided noncompliance between treatment Z and mediator T (i.e., $T^1(k) \in \{0,1\}$ and $T^0(k) = 0$). Additionally, to clearly interpret the causal estimands below, we make the following assumption.

Assumption A1 (Stabilized decision).

There exists a constant k_0 such that $T^z(k)$ is constant for $k_0 \le k$, i.e., $T^z(k) = T^z(k')$ if $k_0 \le k, k'$. Therefore, we simply denote $T^z = T^z(k)$ for $k_0 \le k$.

Assumption A1 states that a student's decision to use the offered ETA becomes constant at or after the minimum valid time k_0. For example, if the minimum valid time is 10 minutes after the test starts, it is assumed that students make their decisions about ETA use by then (as in Figure 1a). Or, if the minimum valid time is the test start time, it indicates that students decide whether to use the offered ETA at or before the start of the test (see Figure 1b). This minimum time should be determined based on domain knowledge. Indeed, we motivate this assumption based on our NAEP data and prior studies (e.g., Mollenkopf, 1960; Ogut et al., 2025; Weinstein & Roediger, 2012), and we provide further discussion of its plausibility in our setting in Section 7.

Under one-sided noncompliance and stabilized decision, we define the causal estimands of interest and discuss identification and estimation strategies for the target causal estimands.

4.1. Estimands

In causal mediation analysis, we decompose the functional total effect (FTE), also referred to as the FATE, into the functional natural direct effect (FNDE) and functional natural indirect effect (FNIE), across the functional domain $\mathcal{K}$. These effects at time k, for $k \in \mathcal{K}$, are formally defined as

$$\begin{aligned}\tau_{\text{TE}}(k) &= E\left\{Y^{1,T^1}(k) - Y^{0,T^0}(k)\right\} \\ &= \underbrace{E\left\{Y^{1,T^1}(k) - Y^{1,T^0}(k)\right\}}_{\text{= functional natural indirect effect}} + \underbrace{E\left\{Y^{1,T^0}(k) - Y^{0,T^0}(k)\right\}}_{\text{= functional natural direct effect}} \\ &= \tau_{\text{NIE}}(k) + \tau_{\text{NDE}}(k).\end{aligned}$$

In words, $\tau_{\text{NDE}}(k)$ measures the expected increase in Y at time k when the treatment changes from $Z = 0$ to $Z = 1$ while holding the mediator constant at the value it would have attained under $Z = 0$. On the other hand, $\tau_{\text{NIE}}(k)$ measures the expected increase in Y at time k when the mediator changes to the value it would have attained under $Z = 0$ while holding the treatment constant at $Z = 1$.[1] These effects can be written as follows under one-sided noncompliance:

$$\tau_{\text{NDE}}(k) = E\left\{Y^{1,0}(k) - Y^{0,0}(k)\right\}, \tag{1}$$

$$\tau_{\text{NIE}}(k) = E\left\{Y^{1,T^1}(k) - Y^{1,0}(k)\right\}. \tag{1}$$

For our ETA example, $\tau_{\text{NDE}}(k)$ means the direct effect of receiving ETA at time k, not mediated by using it, and $\tau_{\text{NIE}}(k)$ means the indirect effect of receiving ETA through its natural use by a student at time k.

Moreover, we are interested in the subgroup effects for the FNDE and FNIE, denoted as S-FNDE and S-FNIE. The S-FNDE and S-FNIE measure the FNDE and FNIE among a subgroup of individuals determined by a subset of pre-treatment covariates V, respectively. These subgroup effects examine effect

[1] We could decompose the total effect into $\tau^*_{\text{NDE}}(k) = E\left\{Y^{1,T^1}(k) - Y^{0,T^1}(k)\right\}$ (i.e., direct effect) and $\tau^*_{\text{NIE}}(k) = E\left\{Y^{0,T^1}(k) - Y^{0,T^0}(k)\right\}$ (i.e., indirect effect). However, in our context, $Y^{0,1}(k)$ is practically not meaningful due to one-sided noncompliance, so we focus on the estimands $\tau_{\text{NDE}}(k)$ and $\tau_{\text{NIE}}(k)$.

https://doi.org/10.1017/psy.2026.10087 Published online by Cambridge University Press

heterogeneity, i.e., how the effects vary across different subgroups. The S-FNDE and S-FNIE are defined as

$$\tau_{\text{NDE}}(k|V) = E\left\{Y^{1,T^0}(k) - Y^{0,T^0}(k) \mid V\right\} = E\left\{Y^{1,0}(k) - Y^{0,0}(k) \mid V\right\}, \tag{3}$$

$$\tau_{\text{NIE}}(k|V) = E\left\{Y^{1,T^1}(k) - Y^{1,T^0}(k) \mid V\right\} = E\left\{Y^{1,T^1}(k) - Y^{1,0}(k) \mid V\right\}. \tag{4}$$

For our ETA example, V can be selected as students' disability status, where $V = 1$ indicates students with disabilities and $V = 0$ indicates those without disabilities. Then, $\tau_{\text{NDE}}(k|V = 1)$ is the direct effect of receiving the ETA at time k among students with disabilities, while $\tau_{\text{NIE}}(k|V = 1)$ is the indirect effect of receiving the ETA via its natural use at time k among students with disabilities. Similarly, $\tau_{\text{NDE}}(k|V = 0)$ and $\tau_{\text{NIE}}(k|V = 0)$ are the corresponding direct and indirect effects, respectively, among students without disabilities.

4.2. Identification

The potential variable $T^z(k)$ (which equals T^z under Assumption A1), i.e., potential ETA use, is inherently derived from a potential functional mediator at time k for subjects with $Z = z$, denoted as $H^z(k)$. Here, $H^z(k)$ represents the amount of time a subject would intend to spend on the test at time k if they received $Z = z$. This variable determines the potential intended ETA use variable at time k, specifically, $T^z(k) = \mathbb{1}\{H^z(k) > c\}$, where c is the standard testing time. Although $H^z(k)$ is only partially observable through H, we introduce the variable to identify the FNDE and FNIE within a functional framework (see Appendix S1 of the Supplementary Material for details on the identification and use of $H^z(k)$). In this article, we focus on the setting of Figure 1a with post-treatment covariates W, due to its added complexities compared to the simpler case of Figure 1b with $W = \{\}$. In the presence of post-treatment covariates, we make the following assumptions to identify the FNDE and FNIE, as well as their subgroup effects (S-FNDE and S-FNIE).

Assumption A2 (Consistency).

(i) If $Z = z$ and $k_0 \le k$, then $T = T^z(k)$ and $\mathbb{1}(k \le H) = \mathbb{1}\{k \le H^z(k)\}$.
(ii) If $Z = z$ and $T = t$ and $k \le H$, then $Y(k) = Y^{z,t}(k)$.
(iii) If $Z = z$, then $W = W^z$.

Assumption A3 (Unconfoundedness). For all $z \in \{0,1\}$:

(i) $Y^{z,t}(\mathcal{K}), H^z(\mathcal{K}), W^z \perp\!\!\!\perp Z \mid X$;
(ii) $Y^{z,t}(\mathcal{K}) \perp\!\!\!\perp H^z(\mathcal{K}) \mid (Z = z, X, W)$.

Assumption A4 (Positivity).

(i) $P_{Z|X}(z \mid x) > 0$ for all z, x.
(ii) $P_{HW|Z,X}(h, w \mid z = 1, x) > 0$ for all x, w and $h \in \mathcal{K}$.
(iii) $P_{HW|Z,X}(h, w \mid z = 0, x) > 0$ for all x, w and $h \in [k_0, c]$.

Assumptions A2–A4 extend the standard identifying assumptions for mediation analysis (Pearl, 2001) to the context of functional outcomes with post-treatment covariates in digital assessments. At a high level, Assumption A2 (Consistency) links observed variables to potential outcomes and mediators; Assumption A3 (Unconfoundedness) states that there are no unmeasured confounders in treatment and mediator assignments, and Assumption A4 (Positivity) ensures sufficient overlap in the distributions of treatment, mediator, and post-treatment covariates between treated units and control units. Note that when post-treatment covariates are absent, one can simply omit or revise parts of the assumptions that include W and W^z. In Appendix S2 of the Supplementary Material, we provide a set of identifying assumptions for this simplified setting.

More specifically, Assumption A2 states that for A2-(i), a subject's potential ETA use under a given treatment at time k, i.e., $T^z(k)$, is identical to their observed mediator for that same treatment T; we

https://doi.org/10.1017/psy.2026.10087 Published online by Cambridge University Press

remark that $T = T^Z = T^Z(k)$ for $k_0 \leq k$ under Assumptions A1 and A2. It also assumes that the indicator of whether $H^z(k)$ (i.e., the potential response time at k) exceeds or equals time k is identical to whether H (i.e., the observed total response time) is larger than or equal to k.[2] Assumption A2-(ii) states that when time k is less than or equal to H (implying that a subject's data is collected at time k), their potential outcome $Y^{z,t}(k)$ is the same as the observed outcome $Y(k)$ if their treatment and mediator were set to $Z = z$ and $T = t$. Likewise, Assumption A2-(iii) states that their potential post-treatment covariate W^z under $Z = z$ is the same as the observed post-treatment covariate W.

Assumption A3-(i) states that within every value of the pre-treatment covariates X, the treatment Z is randomly assigned to subjects, and thus, independent of the potential outcome process $Y^{z,t}(\mathcal{K})$, the potential functional mediator process $H^z(\mathcal{K})$, and the potential post-treatment covariates W^z. In our ETA context, suppose two students have the same pre-treatment covariates, including disability status, English-language learner status, prior math level, and perseverance level. Assumption A3-(i) means that, given these covariates, the ETA receipt status is as good as randomly assigned. Next, Assumption A3-(ii) states that conditional on treatment and the pre- and post-treatment covariates, $H^z(\mathcal{K})$ is independent of $Y^{z,t}(\mathcal{K})$. In particular, it is important to adjust for the post-treatment covariates as well to ensure the independence between $H^z(\mathcal{K})$ and $Y^{z,t}(\mathcal{K})$. To illustrate, consider two students who receive the same treatment (say, $Z = 1$) and share identical pre-treatment covariates (e.g., disability status and prior math level) and post-treatment covariates (e.g., items accessed up to k_0). Assumption A3-(ii) implies that given this information, their potential use of ETA (as determined by their potential total response time) is independent of their potential evolving test scores, and thus, the assignment of the mediator (i.e., the ETA use) is as good as random.

Assumption A4 states that (i) for every value of the covariates, the probability of receiving treatment (i.e., the propensity score) is between 0 and 1, (ii) for every value of the pre- and post-treatment covariates, the density distribution of the observed response time H over $\mathcal{K}$ is non-zero among subjects with $Z = 1$, and (iii) for every value of the pre- and post-treatment covariates, the density distribution of H over $[k_0, c]$ is non-zero among subjects with $Z = 0$. In our ETA example, Assumption A4-(i) means that each eligible student has a real chance of receiving ETA given their pre-treatment covariates, and A4-(ii) and A4-(iii) mean that for each combination of pre-treatment covariates and ETA receipt status, there are students who exhibit all possible values of response time and early-test covariates.

In addition to Assumptions A1–A4, we assume that $Y^{z,t=0}(k) = Y^{z,t=0}(c)$ for $k \geq c$ to reflect that a student's outcome does not change over the extended time period $[c, k_m]$ if they do not use the ETA. In Appendix S1 of the Supplementary Material, we provide formal proofs of identification for the FNDE and FNIE based on these assumptions, building on identification results of Pearl (2001) and Robins (1986). Briefly, the identifying formulas for the FNDE and FNIE depend on the outcome model and the model for the post-treatment variables (i.e., the mediator and the post-treatment confounder), which are defined as

$$m_{zt}(k, w, x) = E\{Y^{z,t}(k) | Z = z, W = w, X = x\}, \quad k \in \mathcal{K} \tag{5}$$

$$g_{zt}(w, x) = P\{T = t, W = w | Z = z, X = x\}, \tag{6}$$

where $(z, t) \in \{(0,0), (1,0), (1,1)\}$. Note that $(z, t) = (0,1)$ is not considered due to one-sided noncompliance. In Appendix S3 of the Supplementary Material, we also outline the identifying assumptions for the general case with a binary static mediator so that readers can adapt them in their specific contexts.

[2] The assumption $\mathbb{1}(k \leq H) = \mathbb{1}\{k \leq H^z(k)\}$ would hold if we make a stronger assumption $H^z(k) = H$. But we do not assume $H^z(k) = H$ because $H^z(k)$ is a time-varying quantity, and it is unrealistic to assume that it is identical to the observed variable H over the entire functional domain. Therefore, we make a weaker and more realistic assumption about the binary relationship between k and $H^z(k)$, which is also sufficient for causal identification.

https://doi.org/10.1017/psy.2026.10087 Published online by Cambridge University Press

4.3. Estimation

To estimate the FNDE and FNIE, as well as their subgroup effects, we use function-on-scalar regression from the functional data analysis literature (Crainiceanu et al., 2024). Let $S = \{W, X\} \in \mathbb{R}^{p+q}$. We estimate the outcome model (5), i.e., $m_{zt}(k,w,x)$, using a GAM in the following additive form:

$$m_{zt}(k,w,x) = m_{zt}(k,s) = \beta_{0zt}(k) + \sum_{r=1}^{R} f_{rzt}(k,s_r). \tag{7}$$

In this model, $\beta_{0zt}(k)$ represents the functional intercept, while $f_{rzt}(k,s_r)$ is its r-th smooth function that depends on a subset of covariates $S_r \subseteq S$ $(r = 1,\ldots,R)$. These smooth functions can vary across covariate domains and the functional domain and may include time-invariant linear terms, linear interaction terms, as well as one-dimensional and multi-dimensional smooth terms.

As an example, consider the following instance of GAMs for the outcome model:

$$m_{zt}(k,s) = \beta_{0zt}(k) + \sum_{j=1}^{p+q} f_{jzt}(s_j) + \sum_{j=1}^{p+q} f_{jzt}(k,s_j). \tag{8}$$

Here, $f_{jzt}(s_j)$ is a time-invariant, univariate smooth function, and $f_{jzt}(k,s_j)$ is a time-varying smooth function. Both are defined for subjects with $Z = z$ and $T = t$ and for each covariate S_j $(j = 1,\ldots,p+q)$.

These smooth functions are parameterized using spline basis expansions. Specifically, for the smooth functions in the outcome model (8), we may take

$$\beta_{0zt}(k) = \sum_{l=1}^{L_{0jzt}} \beta_{0zt,l}\phi_{zt,l}(k), \tag{9}$$

$$f_{jzt}(s_j) = \sum_{l=1}^{L_{1jzt}} \beta_{1jzt,l}\phi_{zt,l}(s_j), \tag{10}$$

$$f_{jzt}(k,s_j) = \sum_{l=1}^{L_{2jzt}} \beta_{2jzt,l}\phi_{zt,l}(k,s_j), \tag{11}$$

where $\{\phi_{zt,l}(\cdot)\}$ are basis functions, and $\{\beta_{(\cdot)jzt,l}\}$ are the corresponding coefficients. Various classes of basis functions can be used, including B-splines, natural cubic splines, and thin plate splines. In this study, we use thin plate splines. For an overview of thin plate splines, see Section 5.5.1 of Wood (2017) and Wood (2003). We select the smoothing parameter, which controls the trade-off between smoothness and model fit, using the generalized cross-validation (GCV) method. Additionally, we estimate the post-treatment variable model (6) using a similar approach, but with a link function (e.g., logit). More details on its estimation are provided in Appendix S4 of the Supplementary Material.

Let us denote estimators of m_{zt} and g_{zt} in (5) and (6) as $\hat{m}_{zt}$ and $\hat{g}_{zt}$, respectively. Using these estimators, we construct estimators for the FNDE and FNIE, and their total effect, FTE, as

$$\hat{\tau}_{\mathrm{NDE}}(k) = \frac{1}{n}\sum_{i=1}^{n}\sum_{w}\{\hat{m}_{10}(k,w,X_i)\hat{g}_{1\cdot}(w,X_i) - \hat{m}_{00}(k,w,X_i)\hat{g}_{0\cdot}(w,X_i)\}, \tag{12}$$

$$\hat{\tau}_{\mathrm{NIE}}(k) = \frac{1}{n}\sum_{i=1}^{n}\sum_{w}\{\hat{m}_{11}(k,w,X_i) - \hat{m}_{10}(k,w,X_i)\}\hat{g}_{11}(w,X_i), \tag{13}$$

$$\hat{\tau}_{\mathrm{TE}}(k) = \hat{\tau}_{\mathrm{NDE}}(k) + \hat{\tau}_{\mathrm{NIE}}(k), \tag{14}$$

where $\hat{g}_{z\cdot}(w,x) = \hat{g}_{z0}(w,x) + \hat{g}_{z1}(w,x)$, which is an estimator for $P(W = w|Z = z, X = x)$. If W is continuous, the summation over w can be replaced with an integral. Likewise, the estimators for the subgroup effects of S-FNDE, S-FNIE, and S-FTE are as follows:

$$\hat{\tau}_{\mathrm{NDE}}(k|v) = \frac{\sum_{i=1}^{n}\mathbb{1}(V_i = v)\sum_{w}\{\hat{m}_{10}(k,w,X_i)\hat{g}_{1\cdot}(w,X_i) - \hat{m}_{00}(k,w,X_i)\hat{g}_{0\cdot}(w,X_i)\}}{\sum_{i=1}^{n}\mathbb{1}(V_i = v)}, \tag{15}$$

https://doi.org/10.1017/psy.2026.10087 Published online by Cambridge University Press

$$\hat{\tau}_{\text{NIE}}(k|v) = \frac{\sum_{i=1}^{n} \mathbb{1}(V_i = v) \sum_w \{\hat{m}_{11}(k,w,X_i) - \hat{m}_{10}(k,w,X_i)\} \hat{g}_{11}(w,X_i)}{\sum_{i=1}^{n} \mathbb{1}(V_i = v)}, \tag{16}$$

$$\hat{\tau}_{\text{TE}}(k|v) = \hat{\tau}_{\text{NDE}}(k|v) + \hat{\tau}_{\text{NIE}}(k|v). \tag{17}$$

For inference, we use nonparametric cluster bootstrapping to obtain standard errors or confidence intervals. Specifically, we resample individuals, rather than time points, to account for unit-specific serial dependence. For each sample, we apply the above GAM estimators, i.e., $\hat{\tau}_{\text{NDE}}(k)$, $\hat{\tau}_{\text{NIE}}(k)$, $\hat{\tau}_{\text{TE}}(k)$, $\hat{\tau}_{\text{NDE}}(k|v)$, $\hat{\tau}_{\text{NIE}}(k|v)$, and $\hat{\tau}_{\text{TE}}(k|v)$, and then compute standard errors or pointwise confidence intervals from the bootstrap distribution (see Appendix S5 of the Supplementary Material for details).

5. Simulation study

5.1. Designs and evaluation

We conducted a simulation study to investigate the finite-sample performance of the proposed estimators. We varied the number of observations n from $\{100,200,500,1,000\}$. We set the standard testing time to $c = 20$, the maximum extended testing time to $k_m = 40$, and the minimum value of the functional domain to $k_0 = 6$. For each subject i, we generated two binary covariates X_{i1} and X_{i2} independently from $\text{Ber}(0.5)$, and the ETA receipt status Z_i from $\text{Ber}(\text{expit}(0.2(X_{i1} + X_{i2} - 1)))$, where $\text{expit}(v) = \exp(v)/\{1 + \exp(v)\}$.

For the outcome, we considered the following two data-generating processes: the linear outcome model and the cubic outcome model. The linear outcome model uses

$$\begin{aligned}
Y_i^{z=0,t}(k) &\sim N(0.25k + 5, 10^2), \quad t = 0,1, \quad k \in [1,20] \\
Y_i^{z=1,t=0}(k) &\sim N(0.25k + 2X_{i1} + 2X_{i2} + 5, 10^2), \quad k \in [1,20] \\
Y_i^{z=1,t=1}(k) &\sim N(0.25k + 0.1(k-5)\mathbb{1}(5 < k) + 2X_{i1} + 2X_{i2} + 5, 10^2), \quad k \in [1,40].
\end{aligned}$$

The cubic outcome model uses, for $\mu_k = 10\{-|k/25 - 1|^3 + 1\} \cdot \mathbb{1}(k \le 25) + 10 \cdot \mathbb{1}(25 < k)$ and $v_k = -10\{-|(k-5)/50 - 1|^3 + 1\}$,

$$\begin{aligned}
Y_i^{z=0,t}(k) &\sim N(\mu_k + 2X_{i1} + 2X_{i2}, 10^2), \quad t = 0,1, \quad k \in [1,20] \\
Y_i^{z=1,t=0}(k) &\sim N(\mu_k + 2X_{i1} + 2X_{i2}, 10^2), \quad k \in [1,20] \\
Y_i^{z=1,t=1}(k) &\sim N(\mu_k + v_k\mathbb{1}(5 < k) + 2X_{i1} + 2X_{i2}, 10^2), \quad k \in [1,40].
\end{aligned}$$

The post-treatment covariate at time $k_0 = 6$ was selected as $W_i^z = \mathbb{1}\{Y_i^{z,T^z}(k = 5) \ge C\}$, where $C = 8.25$ for the linear outcome model and $C = 6.88$ for the cubic outcome model. Thus, W_i^z serves as an indicator of whether a student achieved a high test score (or accessed more items) within the first 5 minutes.

We also generated $H_i^z(k_0)$, i.e., the amount of time that subject i would intend to spend on the test for $Z_i = z$, conditional on the pre-treatment covariates (X_{i1}, X_{i2}) and the post-treatment covariates W_i^z, as follows for $z = 0,1$:

$$P\{H_i^z(k_0) = (10,15,20) \cdot (z+1)\} = \begin{cases} (0.2,0.2,0.6) & \text{if } X_{i1} + X_{i2} = 0 \text{ and } W_i^0 = 0 \\ (0.25,0.25,0.5) & \text{if } X_{i1} + X_{i2} = 0 \text{ and } W_i^0 = 1 \\ (0.3,0.3,0.4) & \text{if } X_{i1} + X_{i2} = 1 \text{ and } W_i^0 = 0 \\ (1/3,1/3,1/3) & \text{if } X_{i1} + X_{i2} = 1 \text{ and } W_i^0 = 1 \\ (1/3,1/3,1/3) & \text{if } X_{i1} + X_{i2} = 2. \end{cases}$$

Roughly speaking, students with higher values of the covariate sum $X_{i1} + X_{i2}$ and $W_i^z = 1$ (i.e., higher early-stage test scores) will likely have shorter actual response times. Given $H_i^z(k_0)$, we defined $H_i^z(k) = H_i^z(k_0)$ for all $k \in \mathcal{K}$. The ETA use variable at time k was determined by $T^z(k) = \mathbb{1}\{H^z(k) > 20\}$. Note that $T^0(k) = 0$ due to one-sided noncompliance in ETA settings. The observed data for subject i were defined

https://doi.org/10.1017/psy.2026.10087 Published online by Cambridge University Press

by $\{X_{i1}, X_{i2}, Z_i, W_i, H_i, T_i, Y_i(k)\}$ for $k \in [6, H_i]$, where $\{H_i, T_i, Y_i(k)\}$ were obtained from $H_i = H_i^{Z_i}(k)$, $T_i = T_i^{Z_i}(k)$, and $Y_i(k) = Y_i^{Z_i, T_i}(k)$.

Using the simulated data, we estimated the FNDE, FNIE, and FTE using the proposed GAM estimators in Section 4.3. We denote these estimators as $\hat{\tau}_{\text{NDE},GAM}$, $\hat{\tau}_{\text{NIE},GAM}$, and $\hat{\tau}_{\text{TE},GAM}$. For implementation, we used the function `gam` from R package `mgcv` (Wood, 2023) to fit the GAMs for both the outcome and post-treatment variable models. Details on the implementation are provided in Appendix S6 of the Supplementary Material. As comparison estimators, we considered two parametric regression-based estimators for the FNDE, FNIE, and FTE, fitting the smooth curves over time through linear and cubic terms. We refer to them as *linear estimators* (denoted as $\hat{\tau}_{\text{NDE},Lin}$, $\hat{\tau}_{\text{NIE},Lin}$, and $\hat{\tau}_{\text{TE},Lin}$) and *cubic estimators* (denoted as $\hat{\tau}_{\text{NDE},Cub}$, $\hat{\tau}_{\text{NIE},Cub}$, and $\hat{\tau}_{\text{TE},Cub}$), respectively (see Appendix S6 of the Supplementary Material for a detailed specification of the coefficients in model (9)). We repeated the simulation 1,000 times.

Based on 1,000 repetitions, we calculated the bias and mean squared error (MSE) of the three estimators, defined as follows for the FNDE:

$$\text{Bias}_{\text{NDE},\star} = \frac{1}{1000}\sum_{j=1}^{1000}\frac{1}{35}\sum_{k=6}^{40}\left\{\hat{\tau}_{\text{NDE},\star}^{(j)}(k) - \tau_{\text{NDE}}(k)\right\}$$

$$\text{MSE}_{\text{NDE},\star} = \frac{1}{1000}\sum_{j=1}^{1000}\frac{1}{35}\sum_{k=6}^{40}\left\{\hat{\tau}_{\text{NDE},\star}^{(j)}(k) - \tau_{\text{NDE}}(k)\right\}^2,$$

where $\star \in \{GAM, Lin, Cub\}$ denotes the estimator type and the superscript $^{(j)}$ represents the simulation repetition index. The bias and MSE for the FNIE and FTE are similarly defined. Smaller bias and MSE values indicate better performance of an estimator.

We also evaluated the performance of the inferential procedure based on the bootstrap. For each k, we calculated the empirical coverage rates of 95% bootstrap confidence intervals for the FNDE based on the proposed GAM method (denoted as $\text{CI}_{\text{NDE},GAM}(k)$) as follows:

$$\text{Coverage}_{\text{NDE},GAM}(k) = \frac{1}{1000}\sum_{s=1}^{1000}\mathbb{1}\left\{\tau_{\text{NDE}}(k) \in \text{CI}_{\text{NDE},GAM}(k)\right\}, \quad k \in \mathcal{K}.$$

The coverage rates of the empirical coverage rates of 95% bootstrap confidence intervals for the FNIE and FTE are similarly defined. The R code for the simulations is available at https://github.com/qkrcks0218/FDA.

5.2. Simulation results

Table 2 summarizes the simulation results under the linear outcome model. In this design, the linear and cubic estimators are correctly specified due to the linear form of the outcome model. However, the cubic estimator includes redundant regressors of higher-order time components. As discussed earlier, the GAM estimator is expected to capture the linear form due to its flexibility and adaptivity. This is confirmed by our finding that all three estimators exhibit negligible bias across all three estimands and sample sizes. In terms of MSE, the linear estimator achieves the smallest value, as it is the correctly specified and parsimonious model. In this design, the linear estimator serves as an oracle estimator. The cubic estimator is less efficient than the linear estimator due to the inclusion of redundant terms. The GAM estimator performs between the linear and cubic estimators, and it consistently shows a smaller MSE than the cubic estimator across all three estimands and sample sizes.

Next, we summarize simulation results under the cubic outcome model in Table 3. In this design, the cubic estimator is still correctly specified, but the linear estimator is mis-specified. Despite the change in the outcome model form, the GAM estimator is expected to capture the true model with no concerns about mis-specification. In line with this expectation, both the GAM and cubic estimators exhibit negligible bias, and the GAM estimator generally shows slightly larger absolute bias. In contrast, the linear estimator shows substantial bias for all three estimands and sample sizes. Regarding MSE, the

https://doi.org/10.1017/psy.2026.10087 Published online by Cambridge University Press

Table 2. Simulation results under the linear outcome model

N	Estimator	FNDE		FNIE		FTE	
		Bias	MSE	Bias	MSE	Bias	MSE
100	GAM	0.052	3.919	−0.046	1.709	0.009	1.810
	Linear	0.056	2.674	−0.044	1.324	0.015	1.147
	Cubic	0.016	6.081	−0.038	2.492	−0.019	2.841
200	GAM	0.003	1.832	0.004	0.756	0.004	0.906
	Linear	0.000	1.123	−0.001	0.522	−0.001	0.556
	Cubic	−0.010	3.025	0.018	1.210	0.007	1.427
500	GAM	−0.012	0.637	0.000	0.242	−0.012	0.327
	Linear	−0.012	0.441	0.001	0.181	−0.011	0.224
	Cubic	−0.013	1.080	−0.006	0.397	−0.019	0.531
1,000	GAM	−0.005	0.340	0.008	0.131	0.003	0.155
	Linear	0.001	0.230	0.003	0.095	0.004	0.104
	Cubic	−0.016	0.576	0.012	0.214	−0.004	0.260

Note: FNDE = functional natural direct effect; FNIE = functional natural indirect effect; FTE = functional total effect; GAM = generalized additive model; Linear = parametric regression with a linear term; Cubic = parametric regression with linear, quadratic, and cubic terms; MSE = mean squared error.

Table 3. Simulation results under the cubic outcome model

N	Estimator	FNDE		FNIE		FTE	
		Bias	MSE	Bias	MSE	Bias	MSE
100	GAM	−0.033	4.063	−0.057	2.088	−0.087	2.094
	Linear	−0.156	2.590	−0.096	2.407	−0.247	2.416
	Cubic	0.006	6.171	0.017	2.634	0.030	2.995
200	GAM	−0.033	2.007	−0.062	0.877	−0.096	1.121
	Linear	−0.141	1.222	−0.110	1.604	−0.251	1.799
	Cubic	−0.022	3.083	0.028	1.231	0.004	1.570
500	GAM	0.015	0.765	−0.059	0.339	−0.044	0.459
	Linear	−0.141	0.471	−0.112	1.201	−0.253	1.433
	Cubic	0.008	1.123	0.024	0.448	0.032	0.606
1,000	GAM	0.027	0.404	−0.069	0.182	−0.041	0.225
	Linear	−0.126	0.247	−0.134	1.113	−0.260	1.302
	Cubic	0.008	0.525	0.001	0.218	0.009	0.268

Note: FNDE = functional natural direct effect; FNIE = functional natural indirect effect; FTE = functional total effect; GAM = generalized additive model; Linear = parametric regression with a linear term; Cubic = parametric regression with linear, quadratic, and cubic terms; MSE = mean squared error.

GAM estimator generally achieves the smallest value, and it outperforms the cubic estimator. Although the GAM estimator exhibits (slightly) larger bias than the cubic estimator, its smaller MSE highlights a superior bias–variance trade-off. In contrast, the linear estimator yields much larger MSE due to model mis-specification.

https://doi.org/10.1017/psy.2026.10087 Published online by Cambridge University Press

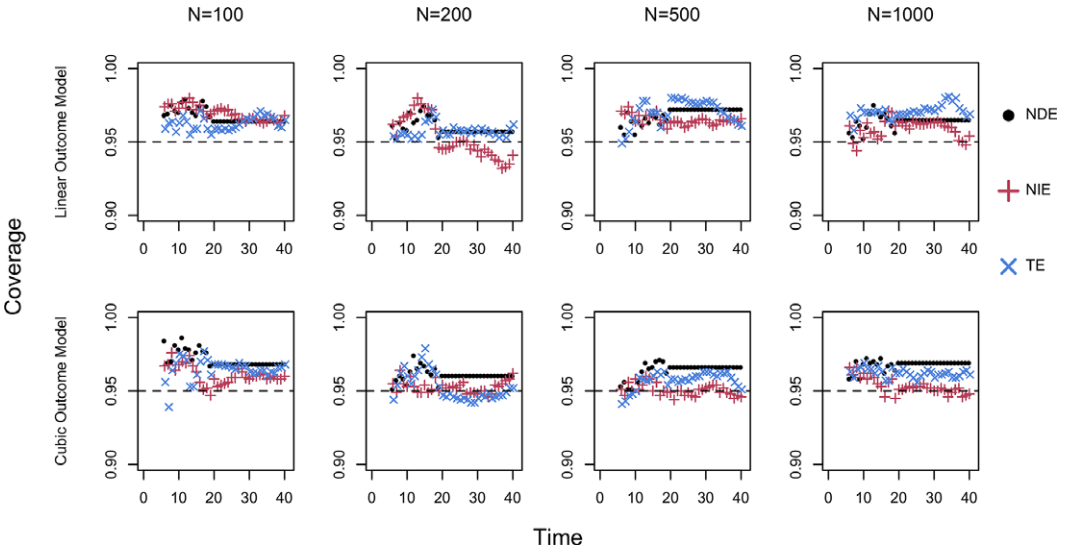


Figure 2. Coverage of 95% bootstrap confidence intervals. The horizontal black line provides a reference for the nominal level 95%; FNDE = functional natural direct effect; FNIE = functional natural indirect effect; FTE = functional total effect.

Figure 2 visually presents the coverage rates under linear and cubic outcome scenarios over the functional domain. In both scenarios, the coverage rates are close to the nominal 95% level across all the estimands and sample sizes. This indicates that the bootstrap procedure provides valid inference.

Overall, our simulation study demonstrates that the GAM estimator performs at least as well as, and often better than, correctly specified parametric models. This emphasizes its capacity to efficiently estimate functional causal effects without concerns about model mis-specification. Additionally, the proposed bootstrap-based method performs well in both outcome models, and it ensures reliable statistical inference with the GAM estimator.

6. Empirical example: ETA

6.1. Data and variables

In our study, we used the 2017 NAEP Grade 8 restricted-use process data for mathematics to estimate the functional total, direct, and indirect effects of receiving ETA. The 2017 NAEP math assessment was administered with multiple test forms, each consisting of two blocks. We used process data from one block, named *M*3, across multiple test forms. Note that test forms were randomly assigned to students, so ignoring form differences should not introduce bias. The selected block contains 15 items, including multiple choice items, selected response items, and constructed response items. The actual items are available on the website: https://www.nationsreportcard.gov/nqt/searchquestions. Within each block, students could revisit and change their answers.

The standard testing duration is 30 minutes, and the extended time adds extra 60 minutes, so the full testing time domain ranges from 1 to 90 minutes. In our data analysis, we set $k_0 = 6$ (i.e., 20% of the standard 30-minute testing time). That is, students' decisions to use ETA are assumed to stabilize after the first 6 minutes, and thus, their early-test responses up to 6 minutes serve as post-treatment covariates. We select $k_0 = 6$ because this initial 20% of the testing period serves as a reasonable window where students form their impressions and decide whether to use ETA. Importantly, we do not claim that $k_0 = 6$ is universally correct. Rather, it is a context-specific choice based on prior literature and testing situations. Using $k_0 = 6$, we incorporate a binary post-treatment covariate about whether the

https://doi.org/10.1017/psy.2026.10087 Published online by Cambridge University Press

number of items accessed by 6 minutes exceeds the sample median. However, if researchers believe that decisions about the use of ETA are made prior to the test, they do not need to include post-treatment covariates, and the setting becomes Figure 1b. Or, if ETA use decisions are believed to stabilize later in the test, say 15 minutes, they may set $k_0 = 15$ and incorporate additional test-taking covariates up to 15 minutes. Additionally, the end of the functional domain k_m was 76 minutes because the range (76, 90] has a relatively small number of students. For these reasons, we used the range [6, 76] as our functional domain of interest.

Our target sample consists of students eligible for ETA, i.e., *students with disabilities* and *English language learners*. We excluded students who spent less than 2 minutes on the test due to extremely low efforts, as well as those who spent more than 76 minutes due to small sample sizes at time k. As a result, our final analysis sample included 4,130 students[3] (99.2% of the eligible sample for ETA in the process data), with about 66% of these students having disabilities, and 33% being English language learners without disabilities.

The functional outcomes of interest are test scores and item access (the number of items a student accessed). We constructed these two outcomes by transforming the raw, unstructured process data into a functional format. The first outcome is a student's evolving test scores on a 15-item math test block over time, and ranges from 0 to 25 points. The second outcome is a student's evolving item access over time, ranging from 0 to 15 items. In this study, we counted a student's access to an item only if they spent more than 3 seconds on it because shorter visits are not considered effortful.

In addition, our treatment variable Z is whether a student received ETA or not, where $Z = 1$ denotes that a student received ETA and $Z = 0$ denotes that a student did not receive it. Our mediator variable T is whether a student used ETA or not, where $T = 1$ denotes that a student used ETA and $T = 0$ denotes that a student did not use it. H is a student's observed total response time, and when H is larger than 32 minutes, $T = 1$; otherwise, $T = 0$. Following prior works (e.g., Lee & Suk, 2025; Suk & Kim, 2025), we use 32 minutes (rather than 30 minutes) as the end of the standard testing period due to interruptions in the typical testing experiences during the NAEP test. We assume one-sided noncompliance between the treatment Z and mediator T, where some ETA recipients did not use ETA, but all non-recipients did not use it. As in prior work, only about 30% of the recipients used ETA, and similar patterns were found among both students with disabilities and English language learners.

Furthermore, we used students' disability status as a moderator V to explore subgroup effects. We also used a set of pre-treatment covariates X, including students' sex, race, disability status, limited English proficiency status, free lunch status, parental education, interest in math, and a proxy for the current math course level. These variables partially explained why some eligible students did not receive or use ETA. For a full list of the variables used in data analysis, see Appendix S7 of the Supplementary Material.

Following our proposed methods discussed in Section 4.3, we used GAM estimators to estimate the FNDE, FNIE, and FTE, as well as their subgroup effects. Following the approach used in the simulation study, we employed the `gam` function from the R package `mgcv` (Wood, 2023) to fit GAMs for both the outcome and post-treatment variable models. In the `gam` function, we selected `knots=18` and `gamma=1.5`, which resulted in no diagnostic issue via the `gam.check` command. See Appendix S6 of the Supplementary Material for details on the implementation. We also conducted nonparametric bootstrapping to estimate confidence intervals for each effect in R software.

As a final note, we did not apply sampling weights and the corresponding jackknife replicate weights provided in the 2017 NAEP data. This allows us to focus on the demonstrations of our methods without introducing additional complexity from the sampling design. However, our results are specific to the analysis sample and should not be generalized to the broader population.

[3]Numbers are rounded to the nearest ten in accordance with NAEP's public reporting policy.

https://doi.org/10.1017/psy.2026.10087 Published online by Cambridge University Press

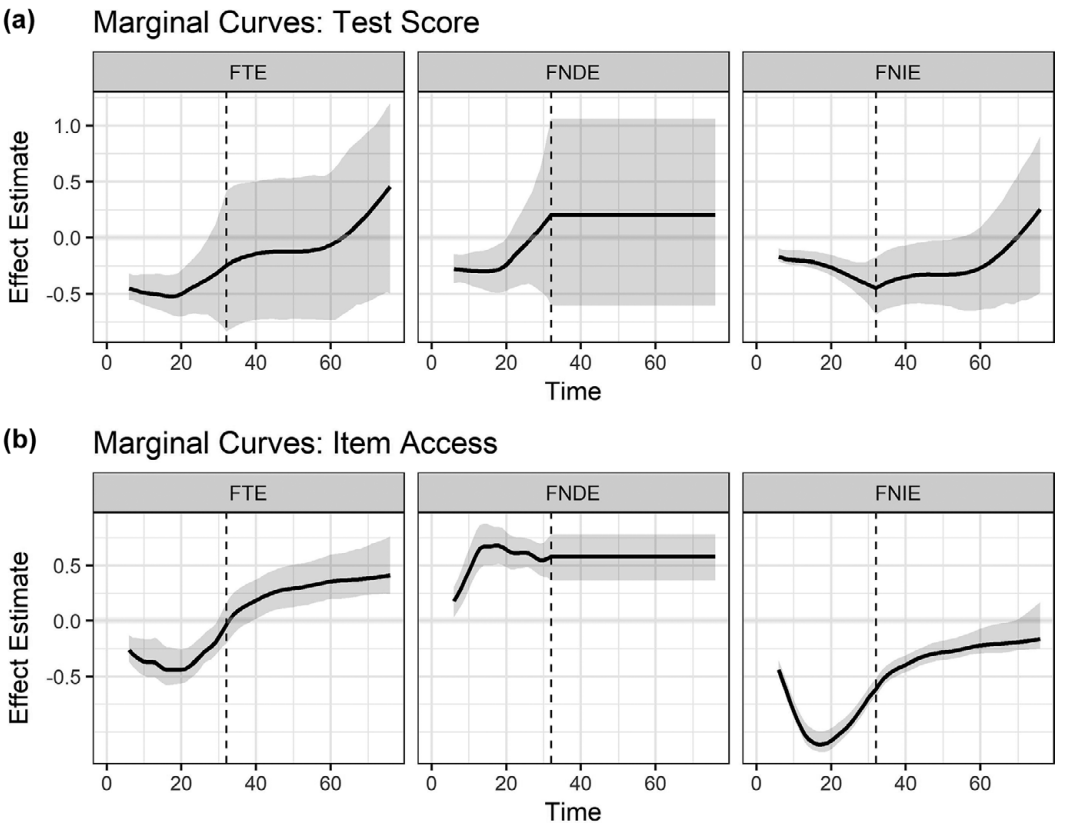


Figure 3. Marginal curves for the functional effects.
Note: Shaded gray area indicates the 95% confidence bands for each effect, a dashed vertical line indicates the end of standard testing time, and a thick gray horizontal line indicates an effect size of zero. FTE = functional total effect; FNDE = functional natural direct effect; FNIE = functional natural indirect effect.
Source: U.S. Department of Education, National Center on Educational Statistics (NCES), 2017 NAEP Grade 8 Mathematics Process Data, Student Features Data File Partial Form and Response Data File.

6.2. Results

Figure 3 summarizes the estimated FTE, FNDE, and FNIE of receiving ETA on two outcomes—test score and item access—over the functional domain ranging from 6 to 76 minutes. The figure includes the 95% confidence bands (shaded gray area) for each effect over time. A dashed vertical line indicates the end of standard testing time (here, 32 minutes), and a thick gray horizontal line indicates an effect size of zero. From Figure 3a, the estimated FTE curve for test scores shows a non-linear trend over time. Initially, the effect is significantly negative and shows a decreasing pattern until around 19 minutes. However, the effect begins to increase after this point, crosses zero near 60 minutes, and continues to rise toward the end of the time period. This pattern suggests that the FTE on test scores improves over time, and a positive effect emerges during the later phase of the extended time period, although it is not significant. After decomposing the FTE into FNDE and FNIE, the direct effect (FNDE) curve starts with a negative effect, but increases sharply between 20 and 32 minutes, remaining constant afterward. This constant effect is expected because the extended time period is inactive for both ETA non-recipients and non-users. In contrast, the indirect effect (FNIE) curve gradually decreases to about −0.4 near 32 minutes. Then, it turns upward, becomes positive after 70 minutes, and continues to increase until the end of the testing period. Overall, during the later phase of the extended time period, the marginal total effect of receiving ETA on test scores (FTE) shows a positive and increasing trend. This total effect is initially

https://doi.org/10.1017/psy.2026.10087 Published online by Cambridge University Press

driven by psychological benefits from receiving ETA on the test scores, and later by the increasing effect of actually using it, although these direct and indirect effects are non-significant throughout the time period.

For item access, shown in Figure 3b, the FTE curve starts with a significantly negative effect and exhibits a decreasing trend until around 18 minutes. After this point, the curve rises rapidly until the end of the standard testing period and then gradually increases during the extended time period. After effect decomposition, the FNDE curve for item access is significantly positive throughout the entire time period, which differs from the results obtained with test scores. It increases rapidly until 15 minutes, slightly decreases by the end of the standard testing period, and then remains constant. This positive direct effect indicates that students who receive ETA but do not use it gain access to items more quickly than those who do not receive ETA. This may be because ETA availability reduces test anxiety or allows students to skim through items initially. In contrast, the FNIE curve declines sharply and reaches its most negative value of −1.7 near 18 minutes. Then, it rises steadily and reaches a level that is not significantly different from zero. This finding indicates that students who use ETA have a slower rate of item access compared to non-users, but eventually catch up on unaccessed items over the extended time period.

To explore heterogeneity of the functional effects, we further examine the S-FTE, S-FNDE, and S-FNIE based on students' disability status (see Figure 4 for the results, where dashed lines represent the 95% confidence bands). For test scores (Figure 4a), students with disabilities (blue curve) show a larger negative effect of FTE compared to those without disabilities (red curve) until 25 minutes. After this point, students with disabilities show a more rapid increase, although the differences in FTE are not significant given the overlapping confidence bands. When focusing on direct and indirect effects, the FNDE is higher for students with disabilities, particularly after 20 minutes, where their curve stabilizes at a higher level compared to those without disabilities. However, these differences are also not significant. In terms of FNIE, students with disabilities show a sharper upward trend after 40 minutes compared to the gradual increase among students without disabilities, but again, the differences are not significant.

For item access (Figure 4b), we observe notable differences between students with and without disabilities. The FTE for item access is more negative for students with disabilities until around 30 minutes, but after this point, the curve becomes positive more quickly and continues to increase until the end. Students without disabilities show a flatter trend after 40 minutes. The differences in FTE between the two groups are not significant. In contrast, we find significant differences in FNDE and FNIE. Specifically, the FNDE curve for students with disabilities rises more sharply and stabilizes at a significantly higher level compared to those without disabilities. This finding indicates that the psychological effect of receiving ETA on item access is significantly greater among students with disabilities. For the FNIE, students with disabilities show a steeper decline initially and show significantly lower effects compared to those without disabilities. However, this negative effect recovers and after around 50 minutes, there are no significant differences between the two groups. This result suggests that although students with disabilities may initially access items at a slower pace, they eventually catch up on unaccessed items within the extended time period.

7. Discussion and conclusions

This article extends the standard causal mediation framework to the context of functional outcomes to identify and estimate the FNDE, FNIE, and FTE, as well as their subgroup effects, using process data from digital assessments. Our framework enables researchers to answer the multi-faceted causal question: "*Why, when, and for whom does a program work?*". Within this framework, we provide the counterfactual definitions of these effects and present nonparametric identification strategies for each effect depending on the presence or absence of post-treatment covariates. To estimate these effects, we construct the GAM-based estimators using function-on-scalar outcome regression and compute their standard errors and confidence intervals via nonparametric bootstrapping. Our simulation study reveals that the GAM estimators perform as well as, and often better than correctly specified parametric

https://doi.org/10.1017/psy.2026.10087 Published online by Cambridge University Press

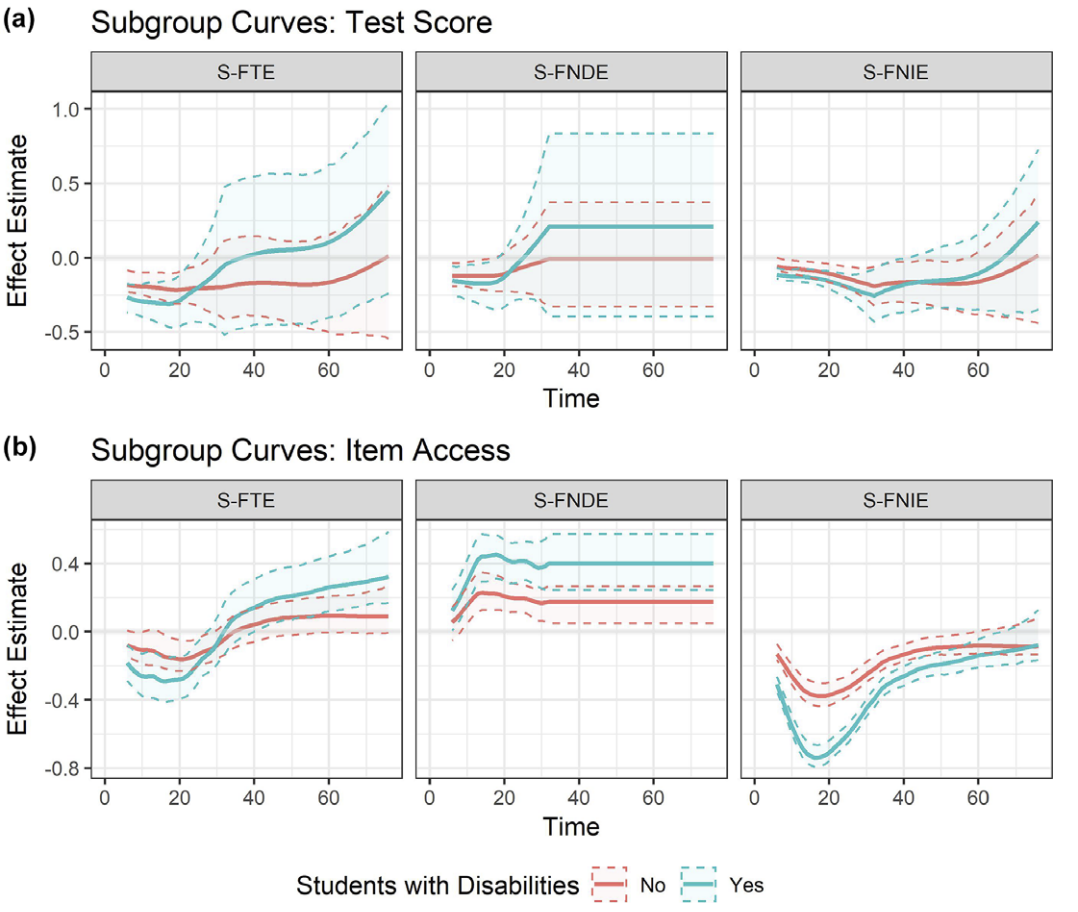


Figure 4. Subgroup curves for the functional effects.
Note: Dashed lines indicate the 95% confidence bands for each effect, and a thick gray horizontal line indicates an effect size of zero. S-FTE = subgroup functional total effect; S-FNDE = subgroup functional natural direct effect; S-FNIE = subgroup functional natural indirect effect.
Source: U.S. Department of Education, National Center on Educational Statistics (NCES), 2017 NAEP Grade 8 Mathematics Process Data, Student Features Data File Partial Form and Response Data File.

methods, with no concerns about model mis-specification. The proposed bootstrap-based method ensures reliable statistical inference with the GAM estimators. Furthermore, we demonstrate our proposed methods in a real-world application on the ETA using NAEP process data. In this analysis, we estimate the FTE, FNDE, and FNIE of receiving ETA for two outcomes—test scores and item access—and also examine their subgroup effects depending on students' disability status. This application uncovers the dynamic causal mechanisms underlying the effects of ETA on functional outcomes and highlights when and for whom each effect works during the testing period.

In our framework, Assumption A1 (stabilized decision) means that students form a plan regarding ETA use before or early in the test. This assumption is crucial for defining the functional effects of interest, but it contrasts with the alternative view that the decision to use ETA evolves continuously until the end of the standard testing time. Empirical evidence supports the stabilized view. In our 2017 NAEP data, more than 25% of the students without disabilities and without ETA experienced time pressure toward the end of the test (Ogut et al., 2025). This suggests a strong need for extra time among our target sample (i.e., students with disabilities and English language learners), but only about 30% of ETA recipients actually used it. This low rate is difficult to explain via a continuously evolving decision process, but aligns well with Assumption A1. In addition, prior research on testing behavior supports

https://doi.org/10.1017/psy.2026.10087 Published online by Cambridge University Press

the plausibility of Assumption A1; students form rapid global impressions of a test's difficulty based on the earliest items (e.g., Anaya et al., 2022; Jackson & Greene, 2014; Weinstein & Roediger, 2012), and they plan test-taking strategies in advance when facing time limits (Mollenkopf, 1960). These findings imply key strategic decisions, including ETA use, are likely determined before or during the early phase of the test.

Moreover, our proposed approach is timely and highly relevant in today's data-rich environment. As digital devices have become widely used, vast amounts of functional data are being generated across various fields. While this rich information is undoubtedly valuable, conventional methods may not be able to handle such complex data. Our approach addresses this challenge by developing a novel causal tool, specifically designed for functional data. While existing causal mediation and moderation methods can reveal the causal mechanisms within subgroups, these methods are still limited in answering crucial questions, such as when or where a program is effective. Our proposed approach extends these existing methods by investigating the causal effects over the functional domain of interest (e.g., time and space) to help uncover more nuanced and detailed causal evidence from complex data. Furthermore, our study tackles another challenge of analyzing time-stamped process data by demonstrating how to generate functional outcomes from such data and examine them with functional data analysis methods. We focus on function-on-scalar regression because of its suitability for functional outcomes, but other functional data methods hold great promise in further advancing the analysis of complex process data from computer-based testing and learning environments.

Based on the findings of this article, we provide limitations of this article and some suggestions for future research. First, in Assumption A1 in Section 4, we assume that test-takers' decisions to use ETA stabilize at a certain time k_0 (e.g., at the start, within the first 20% of the standard testing time). However, in some testing contexts (e.g., high-stakes or highly speeded assessments), the decision to use ETA may not stabilize but instead evolve continuously. To the best of our knowledge, such dynamic decision-making has not been systematically examined, but future research would investigate the stability of these decisions across different testing conditions. Second, if researchers hypothesize a continuously evolving decision regarding ETA use, they should model the dynamic interactions among item access, test scores, and test drop-out (i.e., whether a test-taker exits the test) at each time point. This would require a survival analysis framework that treats test drop-out as a censoring event. In this framework, ETA use should be modeled as a functional mediator rather than a static mediator. Future work would investigate how to extend our framework to the context of functional mediators.

Third, we assume that all the confounders are measured in the observed data. Since observed data may suffer from potential unmeasured confounders, future research would develop sensitivity analysis tools or robust causal methods against unmeasured confounding, for example, by extending proximal causal inference approaches (Tchetgen Tchetgen et al., 2024). Fourth, our estimation methods focus on outcome regression, weighted by quantities derived from the post-treatment confounder model. In our setting, it is possible to develop more robust estimators that satisfy double robustness. Such methods would be consistent if either the treatment assignment (propensity score) model or the outcome model is consistently estimated, thereby increasing robustness to model mis-specification.

Overall, although our approach has some limitations, we believe that our proposed framework and methods contribute to a more nuanced understanding of dynamic causal mechanisms in functional data, ultimately advancing program evaluation efforts in the social sciences.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/psy.2026.10087.

Data availability statement. This article used the restricted-use version of the Grade 8 Mathematics Process Data from the 2017 National Assessment of Educational Progress (NAEP). A restricted-use data license can be obtained through the IES Electronic Application System (https://nces.ed.gov/statprog/instruct.asp).

Acknowledgements. The authors would like to thank three graduate assistants—Ruixuan Li, Yuxin Cai, and Yuxuan Li—for cleaning data and conducting data analysis, and also thank Joseph M. Kush and other audiences who provided comments on

https://doi.org/10.1017/psy.2026.10087 Published online by Cambridge University Press

this work during the 2025 National Council on Measurement in Education (NCME). Additionally, the authors used ChatGPT (GPT-4o/GPT-5) solely for language editing to improve clarity and readability. The tool was not used to generate scientific content, analyze data, or draw conclusions. All ideas and interpretations are the authors' own.

Author contributions. Y.S. and C.P. contributed equally to this manuscript as co-first authors. The order of their names was randomly selected.

Funding statement. The authors received no financial support for the research, authorship, and/or publication of this article.

Competing interests. The authors declare that there are no financial or other competing interests in relation to the work.

References

Anaya, L., Iriberri, N., Rey-Biel, P., & Zamarro, G. (2022). Understanding performance in test taking: The role of question difficulty order. *Economics of Education Review*, *90*, 102293. https://doi.org/10.1016/j.econedurev.2022.102293

Bennett, R. E., Persky, H., Weiss, A. R., & Jenkins, F. (2007). Problem solving in technology-rich environments. (A Report from the NAEP Technology Based Assessment Project, Research and Development Series No. NCES 2007-466). National Center for Education Statistics.

Bergner, Y., & von Davier, A. A. (2018). Process data in NAEP: Past, present, and future. *Journal of Educational and Behavioral Statistics*, *44*(6), 706–732. https://doi.org/10.3102/1076998618784700

Crainiceanu, C. M., Goldsmith, J., Leroux, A., & Cui, E. (2024). *Functional data analysis with R*. (1st ed.) Chapman; Hall/CRC. https://doi.org/10.1201/9781003278726

Ecker, K., de Luna, X., & Schelin, L. (2023). Causal inference with a functional outcome. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *73*(1), 221–240. https://doi.org/10.1093/jrsssc/qlad092

Gao, X., Wang, J., Hu, G., & Sun, J. (2024). Functional causal inference with time-to-event data. *Statistics in Biosciences*. https://doi.org/10.1007/s12561-024-09439-4

Hastie, T., & Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, *1*(3), 297–310. https://doi.org/10.1214/ss/1177013604.

Jackson, A. L., & Greene, R. L. (2014). Impression formation of tests: Retrospective judgments of performance are higher when easier questions come first. *Memory & Cognition*, *42*(8), 1325–1332. https://doi.org/10.3758/s13421-014-0439-5

Jiao, H., He, Q., & Veldkamp, B. P. (2021). Editorial: Process data in educational and psychological measurement. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.793399

Lee, Y., & Suk, Y. (2025). Evidence factors in fuzzy regression discontinuity designs with sequential treatment assignments. *Psychometrika*, *90*(4), 1400–1418. https://doi.org/10.1017/psy.2025.10033

Lovett, B. J., & Leja, A. M. (2013). Students' perceptions of testing accommodations: What we know, what we need to know, and why it matters. *Journal of Applied School Psychology*, *29*(1), 72–89. https://doi.org/10.1080/15377903.2013.751477

Mollenkopf, W. G. (1960). Time limits and the behavior of test takers. *Educational and Psychological Measurement*, *20*(2), 223–230. https://doi.org/10.1177/001316446002000203

Neyman, J. S. (1923). On the application of probability theory to agricultural experiments: Essay on principles. Section 9 (with discussion). *Statistical Science*, *4*, 465–480.

Ogut, B., Circi, R., Huo, H., Hicks, J., & Yin, M. (2025). Running out of time: Leveraging process data to identify students who may benefit from extended time. *International Electronic Journal of Elementary Education*, *17*(2), 253–266. https://doi.org/10.26822/iejee.2025.376

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence* (pp. 411–420). San Francisco: Morgan Kaufmann.

Ramsay, J. O., & Silverman, B. W. (2005). *Functional data analysis*. Springer. https://doi.org/10.1007/b98888

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period application to control of the healthy worker survivor effect. *Mathematical Modelling*, *7*(9), 1393–1512.

Robins, J. M., & Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, *3*(2), 143–155. https://doi.org/10.1097/00001648-199203000-00013

Rosenbaum, P. R. (2010). Evidence factors in observational studies. *Biometrika*, *97*(2), 333–345. https://doi.org/10.1093/biomet/asq019

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1037/h0037350

Suk, Y., & Kim, Y. (2025). Fuzzy regression discontinuity designs with multiple control groups under one-sided noncompliance: Evaluating extended time accommodations. *Journal of Educational and Behavioral Statistics*, *50*(6), 962–984. https://doi.org/10.3102/10769986241268902

Tchetgen Tchetgen, E. J., Ying, A., Cui, Y., Shi, X., & Miao, W. (2024). An introduction to proximal causal inference. *Statistical Science*, *39*(3), 375–390.

Thurlow, M., Lazarus, S. S., Albus, D., & Hodgson, J. (2010). Computer-based testing: Practices and considerations. (Synthesis Report No. 78) National Center on Educational Outcomes, University of Minnesota. Minneapolis, MN.

Wang, J.-L., Chiou, J.-M., & Müller, H.-G. (2016). Functional data analysis. *Annual Review of Statistics and Its Application*, *3*(1), 257–295. https://doi.org/10.1146/annurev-statistics-041715-033624

Wei, X., & Zhang, S. (2023). Extended time accommodation and the academic, behavioral, and psychological outcomes of students with learning disabilities. *Journal of Learning Disabilities*, *57*(4), 242–254. https://doi.org/10.1177/00222194231195624

Weinstein, Y., & Roediger, H. L. (2012). The effect of question order on evaluations of test performance: How does the bias evolve? *Memory & Cognition*, *40*(5), 727–735. https://doi.org/10.3758/s13421-012-0187-3

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *65*(1), 95–114. https://doi.org/10.1111/1467-9868.00374

Wood, S. N. (2017). *Generalized additive models: An introduction with R*. Chapman; Hall/CRC. https://doi.org/10.1201/9781315370279

Wood, S. N. (2023). mgcv [R package version 1.9-0]. R package version 1.9-0. https://cran.r-project.org/web/packages/mgcv/index.html

Yang, S., Lorenzi, E., Papadogeorgou, G., Wojdyla, D. M., Li, F., & Thomas, L. E. (2021). Propensity score weighting for causal subgroup analysis. *Statistics in Medicine*, *40*(19), 4294–4309. https://doi.org/10.1002/sim.9029

Cite this article: Suk, Y. and Park, C. (2026). Causal Mediation and Functional Outcome Analysis with Process Data. *Psychometrika*, 424–444. https://doi.org/10.1017/psy.2026.10087

https://doi.org/10.1017/psy.2026.10087 Published online by Cambridge University Press