How Markets End: Insurance Exhaustion and the Limits of Priority

Dalin Sheng*

May 2026

Abstract

Centralized assignment is often followed by an aftermarket that reallocates vacated and unfilled positions. We model the reliability of this second stage as a design choice. Reliable rescue insures main-round failure and creates its own demand: along an equilibrium exhaustion locus, each additional surviving seat induces another failed-reach claimant after attrition. At a closed-form reliability threshold, the safety tier switches from oversubscription to undersubscription. The best priority rule preserves main-round discipline up to that threshold, and no non-wasteful rule can preserve it beyond: priority reorders the queue but cannot shorten it. Under score priority, welfare has an exact phase-by-phase derivative, and the global optimum is one of three candidates selected by two closed-form comparisons. Residual-seat quality can further discipline reaching, while commitment is required to implement any welfare-improving rescue cap. The model predicts cutoff reversal in supplementary admissions and a calculable demographic transition date at which aftermarket volume changes slope.

JEL classification: C78, D47, D82.

Keywords: market design; matching; clearinghouse; aftermarket; insurance; commitment; waitlists.

1 Introduction

Markets designed to end once often end twice. Each March the National Resident Matching Program announces its match; within days the Supplemental Offer and Acceptance Program

*Southwestern University of Finance and Economics. Email: shengdl@swufe.edu.cn. This work was supported by the National Natural Science Foundation of China (grant number 72503198).

(SOAP) allocates the positions the match left unfilled, a process that in 2012 replaced the informal Scramble.[1] Chinese provinces follow every admission batch with a supplementary round in which unfilled quotas are offered, briefly and by score, to students still unassigned. American colleges spend the summer admitting from waitlists as enrollment deposits melt away. Germany's dialog-oriented DoSV procedure closes with an explicit vacancy-clearing phase. In each case, a main-round assignment is followed by a second process that reallocates seats the first round did not fill or retain.

Matching theory, from Gale and Shapley (1962) to the school-choice program of Abdulkadiroğlu and Sönmez (2003), typically stops at the main-round allocation, even though students, residency applicants, and colleges condition their behavior on the process that follows. Aftermarket rules determine when vacancies are disclosed, whether unclaimed quotas lapse, who may enter, and how claimants are ranked. We ask how these rules feed back into the main round and how a designer should choose them.

Across these institutions, design varies along reliability, priority, and eligibility. Reliability determines how many vacancies survive into the second stage, priority orders claimants when vacancies are scarce, and eligibility determines whether assigned agents may trade up. Increasing reliability can exhaust the claimant buffer that shields the main round from strategic demand. Once that buffer is gone, priority can reorder claimants but cannot preserve discipline; residual-seat quality and commitment then determine what further control is possible.

Reliability affects applications through insurance. A student weighing a selective reach program against a safe alternative trades admission risk against the seat she forgoes, the application-portfolio problem studied by Chade and Smith (2006) and Ali and Shorrer (2025). Rescue after failure lowers the cost of that gamble and induces more students to take it. We call the induced shift *reaching*: strategic over-application in the main round. Evidence that fallback options and waitlist priority encourage more selective applications comes from Akbarpour et al. (2022) and Kuno (2023). Reaching, in turn, adds claimants to the same aftermarket that provides the insurance. Its equilibrium size and composition therefore depend jointly on surviving vacancies and the demand they induce.

The baseline isolates this feedback in a two-tier assignment problem. A unit mass of students with known scores faces a reach tier, selective and noisy in its admissions, worth v_1, and a safety tier, worth $v_2 < v_1$, that admits by score. Each student makes one main-round application; this restriction represents the assignment risk created by list limits, mismatch, and forfeiture rules (Section 2 discusses the choice, and the parallel-list reform analyzed by

[1] In 2010, near the end of the Scramble era, roughly 13,000 applicants competed for about 1,060 unfilled first-year positions (NRMP public data).

Chen and Kesten (2017) supplies its comparative static). After the main round a fraction μ of students melts to outside offers, vacating assigned seats. A fraction ρ of vacated and unfilled seats survives to the aftermarket, where a priority protocol assigns them among students left unmatched, at a delay cost d relative to main-round placement. We compare score priority, a lottery, and buffered priority, which serves safety-tier rejects before failed reach applicants. Reliability ρ summarizes whether vacancy information arrives in time, whether the application window is long enough, and whether unclaimed quotas lapse.

Under score priority, a common rule in supplementary admissions, Theorem 1 characterizes a stable equilibrium that is generically unique and moves through four closed-form branches as reliability rises. Below an erosion onset ρ_0, surviving vacancies are too few to reach the marginal applicant: the aftermarket rehouses unlucky inframarginal students, the threshold stays at the cautious root, and main-round discipline is intact. Between ρ_0 and a second threshold ρ^*, the equilibrium rides a closed-form exhaustion locus along which each additional surviving seat is matched by one additional failed-reach claimant remaining after melt. Insurance supply is exhausted, seat for seat, by the strategic demand it calls forth. At

$$\rho^* = \frac{(1-\mu)\,(1-q_1-q_2)}{\mu\,(q_1+q_2)},$$

the locus reaches the safety tier's subscription boundary. Theorem 1 calls this the *formalization threshold*: beyond it the safety tier is undersubscribed in the main round and fills partly through the aftermarket, while reaching expands toward the full-insurance root. The equilibrium threshold is continuous at an interior ρ^*: only the formula and the composition of supply change, producing a regime kink.

The mechanism is a bang-bang property of marginal insurance. Under score priority the marginal reach applicant stands last in the queue of failed reach applicants, so she is served only when supply covers that entire queue. Along the characterized path her insurance value jumps between zero and $v_2 - d$ where supply meets this demand, and in the middle phase the equilibrium clings to the clearing boundary: that is why the locus is exact.

The formalization threshold is interior, $\rho^* < 1$, precisely when the market is capacity-ample, $\mu(q_1+q_2) > (1-\mu)(1-q_1-q_2)$: melt vacates more seats than the surviving structural shortage. Tight markets remain below the threshold at every feasible reliability, whereas capacity-ample markets cross it once vacancy processing becomes sufficiently reliable. This distinction allows a formalized system such as SOAP to preserve main-round discipline in a tight market even though the same improvement would alter behavior in a slacker market.

Priority design determines how much a designer can gain by changing the queue. Serving safety rejects before failed reach applicants creates a buffer between surviving seats and the strategic margin. Theorem 2 ranks three protocols and establishes the sharp limit of the entire non-wasteful class. A lottery begins eroding discipline at any positive reliability. Score priority holds to ρ_0. Buffered priority holds to ρ^* because total stage-2 demand, failed reach applicants plus safety rejects, telescopes to $(1-\mu)(1-q_1-q_2)$ whenever both classes are nonempty. This is a conservation of claimants: reaching changes who waits, not how many.

Beyond ρ^*, no reordering helps: any non-wasteful protocol must rescue the marginal reach applicant with certainty because supply exceeds total claimant demand; the cautious outcome fails with deviation gain at least $(v_1 - v_2)(v_2 - d)/v_1 > 0$. Priority can reorder the queue; it cannot shorten it. Raising the delay cost d disciplines reaching at the margin; conditional on fully rescuing the marginal applicant into a safety seat, restoring the cautious outcome requires $d \geq v_2$.

The score-priority characterization also yields a welfare rule. Corollary 1 gives the exact welfare derivative in each phase, and Corollary 2 reduces the global optimum to three candidates: complete reliability, the formalization threshold, or one interior throttle point. Two closed-form comparisons select among them by pricing rescue of the displaced against the reaching it subsidizes. Complete reliability wins when delay is sufficiently cheap and again near the upper boundary of the nondegenerate region; capped outcomes arise when the induced-reaching delay loss passes the corresponding dominance test. At the benchmark, ρ^* is the unique optimum; Table B6 maps the other selections and a parameter configuration in which completion always wins.

Residual-seat quality changes the insurance delivered after failure. Proposition 1 derives the depth of the marginal applicant's fallback seat, conserved on one branch and endogenously improving on the other, and shows that greater quality dispersion weakly reduces reaching. The effect stops once reach-tier melt vacancies alone insure the marginal applicant. Theorem 3 then endogenizes the reject buffer under a within-class scramble. Cautious play creates the buffer that sustains caution; aggressive play dissolves it and becomes self-confirming. An analytic mechanism and a numerical sign calculation establish coexistence on an open parameter set.

Commitment and eligibility determine how the market uses reliability. Ex post, serving one more claimant raises realized welfare, so Proposition 2 casts rescue as a time-inconsistent bailout: students anticipate maximal rescue unless the designer can commit to a cap. Section 5 then compares closed eligibility, which serves only the unmatched, with open eligibility, which lets incumbents trade up through vacancy chains. Chains destroy value under common values but can create sorting gains under complementarities.

The model yields observable implications. Along the characterized score-priority path, whenever a safety-tier aftermarket seat goes to a failed reach applicant, its cutoff strictly exceeds the main-round cutoff (Proposition 6). It also identifies the prediction's boundary: once supply goes on to serve safety rejects, the aftermarket cutoff falls below the main-round cutoff. The first pattern is known in Chinese admissions as *daohua*. In official Jiangsu data, after matching regular- and supplementary-round program groups and conservatively excluding institutions whose supplementary plans contain newly added capacity, reversals occur in 138 of 326 comparable groups (42.3 percent) in 2023, 190 of 333 (57.1 percent) in 2024, and 225 of 336 (67.0 percent) in 2025.[2] Because the threshold scales with excess demand, a market with fixed institutional reliability and shrinking cohorts crosses ρ^* at a calculable date. The model predicts a kink: unfilled main-round safety seats appear and aftermarket volume begins rising as the cohort shrinks, although the transition is continuous when $\rho < 1$ (Proposition 7). Aggregate uncertainty weakens insurance when admission and rescue probabilities co-move across years, and signing the effect of information tools on reaching requires a state-contingent application model.

Together, the results give an equilibrium theory in which aftermarket reliability jointly determines main-round behavior, claimant volume, surviving supply, and cutoffs. This interaction produces the four-branch path and the formalization threshold in closed form. The protocol analysis establishes a buffer-capacity bound and a protocol-independent impossibility beyond ρ^*. Within the score-priority environment, the welfare analysis selects the optimal reliability by two closed-form comparisons. The empirical analysis connects the mechanism to cutoff reversals and demographic regime kinks in public admissions data.

1.1 Related literature

Recent work documents later-stage activity around centralized assignment. Kapor et al. (2024) measure the aftermarket surrounding Chilean college admissions and quantify frictions from off-platform options. De Groote et al. (2025) estimate application, acceptance, and waiting decisions in France when centralized admissions coexist with off-platform programs, and Narita (2018) documents how demand evolves after assignment. Applicant behavior also responds in advance: waitlist priority can induce daycare applicants to omit safety options (Kuno, 2023), and valuable outside options can encourage more aggressive school applications (Akbarpour et al., 2022). We model that anticipation as an equilibrium design problem in which reliability determines both surviving supply and the claimant volume generated by reaching.

[2] Author's calculations from Jiangsu Provincial Education Examinations Authority tables (Jiangsu Provincial Education Examinations Authority, 2023–2025); Section 6.1 gives the sample construction, and the replication package gives the official source links, row-level matches, and sensitivity checks. The annual samples are repeated cross-sections, so cross-year differences are descriptive.

The unraveling literature studies markets that move too early. Roth (1984) and Roth and Xing (1994) diagnose unraveling, Niederle and Roth (2003) trace its costs in the gastroenterology market, and Roth and Peranson (1999) analyze the clearinghouse that contains it. Vohra (2025) shows how secondary-market transparency governs early hiring and can eliminate unraveling even when matching remains inefficient. Our aftermarket begins after a centralized assignment and is operated by its designer. Reliability insures main-round failure, so the later market changes the mass of claimants as well as the timing of trade.

Work on two-round and sequential assignment characterizes how staging changes incentives and allocations. It compares sequential and simultaneous systems (Dur and Kesten, 2019), establishes gradual-safety properties for repeated admissions (Haeringer and Iehlé, 2021), identifies when an additional stage improves welfare (Doğan and Yenmez, 2023), and studies systems in which public and private schools admit at different dates (Andersson et al., 2024). Gong and Liang (2025) combine theory and experiment to evaluate real-time feedback in a dynamic admissions mechanism. In our model the flow into the later stage is endogenous: reliability changes both surviving seats and failed-reach demand. The threshold ρ^* locates where this feedback exhausts the claimant buffer, and Corollary 1 values the resulting allocation and delay.

Dynamic reassignment work studies preference discovery through platform offers (Grenet et al., 2022), seat reassignment after late cancellations (Feigenbaum et al., 2020), and trading-chain or priority reforms when preferred positions become available over time (Huitfeldt et al., 2024). Related work characterizes rationing in overloaded waitlists (Leshno, 2022) and the information a waitlist should reveal (Ashlagi et al., 2025). We connect these technologies to the initial reach–safety choice and characterize how reliability and eligibility govern entry into reassignment and the vacancy chains that follow.

Queue-design papers study stopping and allocation on the kidney waitlist (Agarwal et al., 2021), lotteries and waitlists for housing (Arnosti and Shi, 2020), deferral rights under risk aversion and impatience (Schummer, 2021), and the equity–efficiency trade-off of extreme waitlist policies (Nikzad and Strack, 2024). More recent work designs waitlists with an outside option (Afacan and Cumbul, 2025), jointly chooses queue entry, exit, priority, and information (Che and Tercieux, 2026), and derives simple priority structures by coarsening scores (Çelebi and Flynn, 2022). Our priority theorem makes claimant mass responsive to anticipated rescue. Buffered priority preserves discipline through ρ^*, while every non-wasteful ordering fails beyond that bound.

Coexistence of centralized and decentralized channels is studied in static settings by Manjunath and Turhan (2016), Doğan and Yenmez (2019), and Ekmekci and Yenmez (2019). Timing separates an aftermarket from a parallel market: it opens only after main-round outcomes are realized, which makes it insurance and lets it reshape the round it follows. On the supply side, Ashlagi and Roth (2014) and Agarwal et al. (2019) show hospitals withholding kidney-exchange pairs for internal matching; Online Appendix Section A.6 transposes that participation margin to admission platforms.

Arnosti (2015) directly studies how an aftermarket affects the quantity and quality of matches formed by a clearinghouse when submitted lists are short. Auction resale supplies a parallel design logic: Zheng (2002) designs optimal auctions anticipating resale, Haile (2003) characterizes bidding when resale opportunities follow the auction, and Garratt and Tröger (2006) show how resale invites speculation. Trade following a primary allocation can therefore reshape behavior in that allocation. We make reliability, priority, and eligibility joint design choices and derive the resulting capacity exhaustion. Avery and Pathak (2021) show that Tiebout exit constrains what school choice can redistribute; Section 6 Online Appendix Section A.6 develops the analogous constraint from off-platform participation.

The analysis combines the continuum-and-cutoff apparatus of Azevedo and Leshno (2016) with the reach–safety portfolio logic of Chade and Smith (2006) and Ali and Shorrer (2025). The application problem under admission uncertainty builds on Avery and Levin (2010) and Che and Koh (2016), while large-market strategic simplicity in the spirit of Kojima and Pathak (2009) and Azevedo and Budish (2019) motivates the continuum treatment. Doval (2022) develops dynamic stability for matching without a designed second stage. The commitment analysis descends from Kydland and Prescott (1977) through the bailout model of Farhi and Tirole (2012), and the vacancy-chain arithmetic of Section 5 builds on Blum et al. (1997).

Section 2 sets out the model. Sections 3–5 characterize equilibrium, welfare, priority, and eligibility. Section 6 develops empirical implications and institutional evidence. The Online Appendix contains the proofs, a supply-side extension, and numerical verification against directly computed equilibria.

2 Model

The model links one application decision to a later allocation stage: a student chooses between a selective application and a safe one, knowing that the aftermarket partially insures failure. The clearinghouse assigns two tiers of seats in the main round, attrition creates vacancies, and the aftermarket offers vacated and unfilled seats to students left unmatched. Anticipated service in the last step changes behavior in the first.

2.1 Environment

A unit mass of students is indexed by a score $s \sim U[0,1]$, which each student knows. Seats come in two tiers. The *reach tier* (tier 1) has capacity $q_1 \in (0,1)$ and delivers value v_1 per seat; the *safety tier* (tier 2) has capacity $q_2 \in (0,1)$ and delivers v_2, with $v_1 > v_2 > 0$. Preferences are vertical: every student ranks a reach seat above a safety seat above the outside option, whose value we normalize to zero. Between the two assignment stages a student may receive an outside offer worth w: a job, or an offer from a school outside the clearinghouse. A seat obtained in the aftermarket at tier j is worth $v_j - d$, where d is the *delay cost* of late placement.

Assumption 1 (Payoffs)**.** $0 < d < v_2 < v_1$, and outside offers satisfy $w > v_1$.

Assumption 1 serves two roles: $d < v_2$ gives a late safety seat positive value, so rescue is worth having, and $w > v_1$ makes attrition independent of the student's assignment.

2.2 Timing

The market unfolds in three steps.

1 (main round). Each student applies to exactly one tier. The reach tier admits by a noisy index; the safety tier admits by score.

Melt. Each student independently receives an outside offer with probability μ and exits. Assigned leavers vacate their seats; unassigned leavers exit the applicant pool.

2 (aftermarket). Each vacated or unfilled seat survives with probability ρ. Surviving reach-tier seats are assigned before surviving safety-tier seats, and claimants are ordered by a priority protocol π. A student placed at tier j receives $v_j - d$.

In stage 1, applying to the reach tier—*reaching*—is risky. Admission is governed by the index $e = s+\eta$, where $\eta \sim U[-\varepsilon, \varepsilon]$ i.i.d. is admission noise: grading error, subjective review, reassignment risk across programs; we take $\varepsilon > 0$. If reach applications exceed q_1, seats go to the highest indices down to a cutoff κ, so a reach applicant with score s is admitted with probability

$$p_1(s) \;=\; \min\bigl\{1, \max\{0, (s-\kappa+\varepsilon)/(2\varepsilon)\}\bigr\},$$

strictly increasing on the noise band. The safety tier admits deterministically by score: if its applicant mass is at most q_2, it admits everyone and the residual seats go unfilled; if it is oversubscribed, it admits the top q_2 by score and rejects the rest. The noisy tier is where students gamble; the deterministic tier is where they are rationed.

Because $w > v_1$, an outside offer is accepted regardless of assignment, so a fraction $\mu \in (0,1)$ of every group exits: the *melt rate* is invariant to application behavior. A vacated or unfilled seat survives with probability $\rho \in [0,1]$, the aftermarket's *reliability*: whether a vacancy is detected and posted in time, and whether the rules permit it to be filled before quotas lapse. Low ρ describes an informal scramble in which most vacancies lapse; ρ near one describes a formalized second round in which almost all survive. A student served in stage 2 is *rescued*. Baseline eligibility is closed: only students left unmatched by the main round may claim aftermarket seats, as in SOAP and the supplementary round. Section 5 extends eligibility to incumbents.

Given the vertical structure, we study threshold behavior: under the strategy σ_{s^*}, a student applies to the reach tier if and only if $s \geq s^*$. (Section 3 proves single crossing and existence of a threshold best-response selection; where a payoff-neutral interval supports others, ACE selects the threshold one.) All aftermarket flows then have closed forms. Failed reach applicants who survive melt number

$$D_f \;=\; (1-\mu)\max\{0,\; 1-s^*-q_1\},$$

and surviving safety-tier rejects number

$$D_r \;=\; (1-\mu)\max\{0,\; s^*-q_2\}.$$

On the supply side, when the safety tier is oversubscribed the only vacancies are melt vacancies, $S = \rho\,\mu\,(q_1+q_2)$; when it is undersubscribed, unfilled seats add to them, $S = \rho\,[(q_2-s^*)+\mu(q_1+s^*)]$. We call the market *capacity-ample* if $\mu(q_1+q_2) > (1-\mu)(1-q_1-q_2)$, so melt vacancies exceed the surviving structural shortage. In a capacity-ample market, sufficiently reliable supply can cover the entire post-main-round claimant pool.

2.3 Aftermarket protocols

The designer chooses reliability ρ and the priority protocol π that orders claimants when surviving supply is short. Score priority π^{score} serves unmatched claimants in order of s, as in the Chinese supplementary round. The lottery π^{lot} serves claimants in uniform random order, representing an unorganized scramble. Buffered priority π^{buf} serves safety-tier rejects first (the *buffer*) and failed reach applicants last, ranked by score within class, so the marginal reach applicant is the residual claimant of the market. Within any claimant order, reach vacancies are used before safety vacancies. Section 4 defines the non-wastefulness condition used to compare these protocols.

2.4 Equilibrium

The equilibrium object is the marginal student's comparison between a partially insured gamble and a sure seat. To state it, fix a threshold s^* with the reach tier oversubscribed and let $\kappa(s^*)$ denote the market-clearing cutoff, whose existence and uniqueness Section 3 establishes. The *marginal admission rate*

$$p_m(s^*) \;=\; p_1\big(s^*; \kappa(s^*)\big)$$

is the admission probability of the marginal reach applicant herself, recognizing that the pool and the cutoff move with the threshold. Raising s^* weakly raises p_m: the fewer students gamble, the safer the gamble becomes for the one at the margin. Under bounded admission noise, however, p_m can equal zero on an initial interval even while the reach tier is oversubscribed. The regularity condition therefore imposes strictness only where the marginal admission probability is interior, the region the equilibrium proofs use.

Assumption 2 (Regularity)**.** The marginal admission rate $p_m(s^*)$ is continuous and weakly increasing wherever the reach tier is oversubscribed, and strictly increasing wherever $0 < p_m(s^*) < 1$. At every candidate threshold used by a numbered result, the lower edge of the reach admission band lies weakly above the safety cutoff:

$$\kappa(s^*) - \varepsilon \;\geq\; (s^* - q_2)^+.$$

Online Appendix B verifies both conditions over the computed range.

The band condition makes the fallback genuinely safe: any score with an interior reach chance clears the safety cutoff.

The aftermarket enters the application decision through a single statistic. Let $W_2^m(s^*)$ denote the expected stage-2 value of the marginal reach applicant conditional on failing at the reach tier and staying in the market: the insurance the aftermarket offers her, determined by ρ, the protocol π, and the flows (D_f, D_r, S). Conditional on not melting, reaching yields $p_m v_1 + (1 - p_m) W_2^m$, while the safety application yields v_2, since in the relevant range the marginal student is admitted to the safety tier with certainty. The *marginal incentive* is therefore

$$g(s^*) \;=\; p_m(s^*)\big(v_1 - W_2^m(s^*)\big) - \big(v_2 - W_2^m(s^*)\big).$$

When $g(s^*) > 0$ the marginal student strictly prefers to reach and the threshold falls; when $g(s^*) < 0$ she prefers safety and the threshold rises. Because W_2^m enters g with coefficient $1 - p_m > 0$, greater insurance after failure makes reaching more attractive.

Definition 1 (Aftermarket-consistent equilibrium)**.** An *aftermarket-consistent equilibrium* (ACE) is a triple consisting of a threshold s^*, a reach-tier cutoff κ, and a profile of stage-2 service probabilities such that (i) given the aggregates, every student's application is optimal, with the marginal student indifferent at s^*; (ii) κ clears the reach tier; and (iii) the stage-2 service probabilities are those induced by the protocol π and the flows generated by (i) and (ii).

An ACE is *stable* if g crosses zero from below at s^*: after a small perturbation of the threshold in either direction, the marginal applicant's incentive pushes it back. One convention completes the definition. Where surviving supply exactly equals the demand of the class containing the marginal applicant, W_2^m jumps and exact indifference can fail. Because the boundary score has zero mass, its service probability can be selected without changing aggregate feasibility; the *marginal rationing convention* selects the probability that makes this type indifferent whenever the two one-sided incentives straddle zero. The equilibrium of Theorem 1 spends an interval of reliabilities on this locus.

Utilitarian welfare W is the sum of seat values net of stage-2 delay costs. Outside offers add the constant μw and are omitted. Because values are vertical and common within a tier, welfare depends only on the masses assigned at each tier and date; admission noise affects welfare only through the equilibrium threshold.

2.5 Discussion of the assumptions

One application. A single application represents main-round assignment risk. Clearinghouses create such risk through list-length caps, costly applications, and forfeiture rules (in China, a student who declines reassignment to an undersubscribed program is dropped from the entire batch). The mechanism requires a positive probability of remaining unmatched that responds to the student's application. With unrestricted lists and no forfeiture, the main round would absorb reach failure, D_f would vanish, and the aftermarket would process melt alone. The Chinese parallel-list reform, staggered across provinces from 2008 to 2012 (Chen and Kesten, 2017), provides variation toward longer lists; Section 6 states the corresponding prediction.

Uniform noise. Admission noise represents the evaluation uncertainty documented in decentralized admissions (Che and Koh, 2016). Uniformity makes p_1 piecewise linear and the clearing condition closed form. The phase characterization uses the distribution through Assumption 2; Online Appendix B verifies that assumption for the uniform case.[3] At $\varepsilon =$

[3] For other noise distributions, monotonicity of p_m is maintained directly as part of Assumption 2.

0 the marginal applicant's admission probability jumps from zero to one, eliminating the interior reach–safety trade-off.

Exogenous melt. Because $w > v_1$, exit is uncorrelated with assignment, and vacancy supply is the fixed flow $\mu(q_1 + q_2)$. This isolates the demand side of the aftermarket—the choice of who gambles in anticipation of rescue—from supply-side endogeneity. If attrition were selective, S would depend on the composition of the assigned and the exhaustion locus of Theorem 1 would no longer be linear in ρ, though the insurance logic survives. Section 5 relaxes the one-shot vacancy structure.

Vertical tiers. Two tiers with common values isolate the portfolio margin of application choice (Chade and Smith, 2006) while keeping the continuum-and-cutoffs apparatus of Azevedo and Leshno (2016) available in its simplest form. One tier would eliminate portfolio choice altogether; horizontal preference heterogeneity would multiply cutoffs without changing the accounting identities that define exhaustion. Proposition 1 shows how within-tier dispersion weakens the homogeneous-seat impossibility and where reach-tier melt prevents it from restoring caution.

The baseline holds transfers, horizontal preference heterogeneity, and school behavior fixed, so changes in application incentives and claimant volume can be traced to reliability. Section 3 characterizes equilibrium under score priority and locates the erosion onset ρ_0 and formalization threshold ρ^*.

3 Insurance Exhaustion and the Formalization Threshold

Under score priority, aftermarket reliability determines both the supply of rescued seats and the application behavior that generates claimants. Two lemmas reduce equilibrium to a single-crossing problem. Theorem 1 traces the equilibrium path, and Corollaries 1 and 2 characterize the welfare choice between capping and completing the market.

3.1 Threshold play and market clearing

Because the reach-minus-safety payoff difference is monotone in score, best responses admit a threshold selection and aggregate play collapses to s^*.

Lemma 1 (Threshold best responses). *Under score priority, a global lottery, or buffered priority with either score or lottery order within the failed-reach class, and at aggregates satisfying the separation clause of Assumption 2, the reach-minus-safety payoff difference is* *nondecreasing in score. Consequently the best-response correspondence admits a threshold selection: there exists $\hat{s}$ such that the selected action is reach if and only if $s \geq \hat{s}$. Absent a positive-length payoff-neutral interval, every pure best-response selection is threshold up to the zero-mass boundary type.*

Single crossing does the work. The admission probability $p_1(s)$ rises with the score, and under score priority so does the continuation value of failure; the reach gamble thus improves with s on both counts while the safety payoff is constant over the relevant range. The payoff difference can be zero on an interval at a neutrality knife edge, but otherwise crosses once (Online Appendix A). ACE uses the monotone threshold selection: students in $[s^*, 1]$ reach, students in $[0, s^*)$ play safe.

Given threshold play, the reach tier's cutoff is mechanical.

Lemma 2 (Market clearing)**.** *If the reach tier is oversubscribed, the cutoff $\kappa(s^*)$ exists and is unique, and $p_m(s^*) = p_1(s^*; \kappa(s^*))$ is continuous.*

Existence and uniqueness follow because admitted mass is continuous and strictly decreasing in the cutoff, as in Azevedo and Leshno (2016). Equilibrium depends on $p_m(s^*)$, the marginal applicant's admission rate, which Assumption 2 makes weakly increasing globally and strictly increasing on the interior-probability range containing the benchmark roots. Section 2 defines equilibrium as a stable crossing of $g(s^*) = p_m(s^*)\big(v_1 - W_2^m(s^*)\big) - \big(v_2 - W_2^m(s^*)\big)$. We next derive how the aftermarket determines $W_2^m(s^*)$.

3.2 Two benchmark roots

Two polar aftermarkets bracket the equilibrium. In the first, the marginal applicant receives nothing: $W_2^m = 0$. Reaching is then a pure portfolio choice in the spirit of Chade and Smith (2006), a gamble worth v_1 with probability p_m against a certain v_2. The marginal applicant is indifferent when $p_m(s_c^*) = v_2/v_1$, which defines the *cautious root* s_c^*. The required odds are demanding whenever the tiers are close substitutes: at the parameters of Online Appendix B, where $v_2/v_1 = 0.8$, a student reaches only if she expects to win four times out of five.

In the second, failure is fully insured: the marginal failed applicant is rescued into a safety seat with certainty, at delayed value $v_2 - d$. The gamble risks only the delay, and indifference requires $p_m(s_a^*) = d/(v_1 - v_2 + d)$, which defines the *full-insurance root* s_a^*. Because $d < v_2$, the full-insurance odds are strictly below the cautious odds, so $s_a^* < s_c^*$: insured students reach more. The gap is wide: at the same parameters, the required winning probability falls from 0.80 to 0.20. Cheap insurance makes long-shot gambles worthwhile.

Assumption 3 (Nondegenerate configuration)**.** $0 < s_a^* < q_2 < s_c^* < 1 - q_1$ and

$$\mu q_1 < (1-\mu)(1 - s_a^* - q_1).$$

Thus both claimant classes are nonempty at the cautious root and, along the score-priority path, reach-tier melt vacancies alone never cover failed-reach demand, so the marginal rescue, when it occurs, is a safety-tier seat.

We maintain the payoff, regularity, and nondegenerate assumptions throughout Section 3 and for the baseline priority and commitment results in Section 4. Proposition 1 uses its own stated conditions.

Both roots are endogenous. Whether the marginal applicant is rescued depends on the mass of failed applicants competing for surviving seats, which depends on the threshold and hence on anticipated rescue. Theorem 1 resolves this fixed point.

3.3 Exhaustion and the threshold

The fixed point turns on aftermarket accounting. Under score priority the marginal failed applicant has the lowest score in her class, so she is served only if the entire class is. Along the equilibrium path, Assumption 3 ensures that reach-tier vacancies alone never cover that class; her insurance is therefore bang-bang, worth $v_2 - d$ when $S \geq D_f$ and zero when $S < D_f$. Off the path, sufficiently cautious play can instead give her a reach seat; the proof shows that this third branch creates no additional equilibrium.

Theorem 1 (Insurance exhaustion and the formalization threshold)**.** *Under score priority, a stable equilibrium exists and is generically unique. Let $a = 1 - \mu$, $q = q_1 + q_2$, let s_c^* solve $p_m(s_c^*) = v_2/v_1$ and s_a^* solve $p_m(s_a^*) = d/(v_1 - v_2 + d)$, and define*

$$\rho_0 = \frac{a(1 - q_1 - s_c^*)}{\mu q}, \qquad \rho^* = \frac{a(1-q)}{\mu q}.$$

If $\rho^* < 1$, also define

$$\rho_a = \frac{a(1 - q_1 - s_a^*)}{q_2 + \mu q_1 - a s_a^*},$$

the reliability at which the marginal applicant first receives full insurance. Away from the *knife edge $\rho^* = 1$, the equilibrium threshold is*

$$s^*(\rho) \;=\; \begin{cases} s_c^* & \rho \le \rho_0 \quad \textit{(cautious phase)} \\ 1 - q_1 - \dfrac{\rho\mu q}{a} & \rho_0 < \rho \le \min\{\rho^*, 1\} \quad \textit{(oversubscribed exhaustion)} \\ \dfrac{a(1-q_1) - \rho(q_2 + \mu q_1)}{a(1-\rho)} & \rho^* < \rho < \rho_a \quad \textit{(undersubscribed exhaustion)} \\ s_a^* & \rho \ge \rho_a \quad \textit{(full insurance).} \end{cases}$$

The last two branches apply only when $\rho^* < 1$. In the oversubscribed exhaustion phase the equilibrium is pinned by $S = D_f$: every additional surviving seat is matched by one additional failed-reach claimant remaining after melt. The formalization threshold is interior ($\rho^* < 1$) if and only if the market is capacity-ample: $\mu(q_1 + q_2) > (1-\mu)(1-q_1-q_2)$.

We call ρ_0 the *erosion onset* and the middle branch the *exhaustion locus* $s^*_{\text{pin}}(\rho)$. Below ρ_0, surviving seats do not reach the marginal applicant at cautious play, so the aftermarket processes melt without insuring the strategic margin. At ρ_0, supply catches up with cautious failed-reach demand and the equilibrium moves onto $S = D_f$. More cautious play would leave seats to insure the margin and pull the threshold down; more aggressive play would make rescue fail and push it back up. The identity $\rho\mu(q_1+q_2) = (1-\mu)(1-s^*-q_1)$ implies that one additional surviving seat generates one additional failed-reach claimant after melt, or $1/(1-\mu)$ additional reach applications before melt. This is aggregate moral hazard: anticipated rescue induces the reaching that claims the added supply.

Aftermarket volume is an equilibrium outcome on the exhaustion locus. Because reaching expands with rescue, higher vacancy survival need not shorten the supplementary queue; queue length alone can understate the effect of formalization. Composition is informative: along the locus, the added claimants are failed-reach applicants who would have chosen safety under a less reliable aftermarket. Section 6 develops this measurement implication.

The oversubscribed exhaustion phase consumes the safety tier from above and cannot last. Each recruited gambler would otherwise have claimed a safety seat in the main round, so as ρ rises the mass of safety applicants falls toward capacity q_2. At $s^*_{\text{pin}}(\rho^*) = q_2$, safety-tier rejections vanish. Past ρ^* the safety tier is undersubscribed in stage 1 and supply switches to $S = \rho[(q_2 - s^*) + \mu(q_1 + s^*)]$. The equilibrium initially remains on $S = D_f$, now following the third line of the theorem; only at ρ_a does supply become slack at s_a^* and insurance become complete. The threshold is continuous at an interior ρ^*, but its formula and the composition of supply change there. The crossing is a regime kink, not generically a jump: before ρ^* the aftermarket repairs a main round that fills the safety tier; after it, the aftermarket becomes

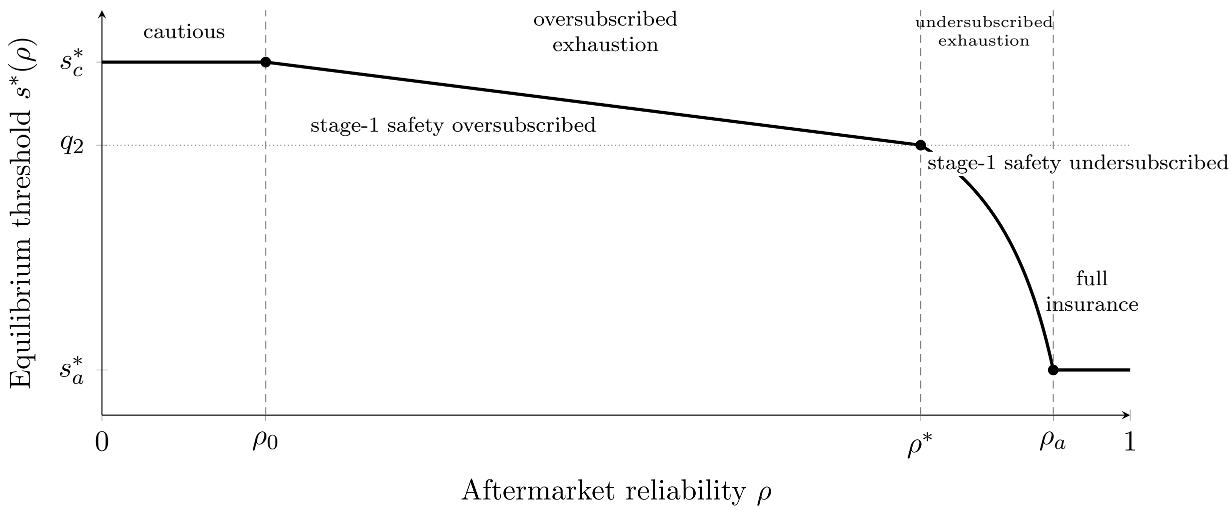

Figure 1: The four-branch equilibrium path under score priority. The figure plots Theorem 1 at the benchmark parameters $(q_1, q_2, \mu, v_1, v_2, d, \varepsilon) = (.20, .75, .062, 1, .80, .05, .25)$. The dotted line marks safety capacity q_2. At ρ^* the safety tier changes from oversubscription to undersubscription; $s^*(\rho)$ remains continuous while its formula and the source of aftermarket supply change.

one of the channels through which that tier fills.

Whether feasible reliability reaches ρ^* depends on primitives: the threshold is interior exactly when $\mu(q_1 + q_2) > (1 - \mu)(1 - q_1 - q_2)$, that is, when seats vacated by melt exceed the surviving structural shortage. When structural shortage dominates melt, $\rho^* > 1$ and formalization cannot cross the threshold. This case is consistent with the replacement of the NRMP Scramble by SOAP without a visible change in main-round behavior (Roth and Peranson, 1999). Shrinking cohorts and expanded capacity move Chinese admission batches in the opposite direction (Chen and Kesten, 2017). Proposition 7 converts these comparative statics into a dated prediction.

The proof, in Online Appendix A, rests on three observations: under score priority W_2^m moves through an ordered set of bang-bang branches; within each branch g is increasing on the equilibrium-relevant interval, where p_m is interior; and each coverage boundary is crossed upward as s^* rises. The off-path reach-seat branch lies above the safety-seat branch and cannot create another crossing. The unique upcrossing is therefore either a benchmark root or the boundary $S = D_f$. The marginal-rationing convention supports the latter without altering aggregate flows.

3.4 Welfare comparison under score priority

Theorem 1 maps reliability into allocations; the welfare question is which reliability to choose. In the vertical baseline there is no selection margin: every reach seat is worth v_1 and every safety seat v_2, so welfare depends only on assignment masses, tier composition, and delay.

Corollary 1 (Welfare and optimal reliability)**.** *Let $a = 1 - \mu$, $q = q_1 + q_2$, $C = q - a$, and suppose $\rho^* < 1$. Then:*

$$\frac{dW}{d\rho} = \begin{cases} \mu\big[q_1(v_1 - d) + q_2(v_2 - d)\big] > 0, & 0 \le \rho < \rho^*, \\ \mu q_1(v_1 - v_2) - \dfrac{Cd}{(1-\rho)^2}, & \rho^* < \rho < \rho_a, \\ \mu q_1(v_1 - v_2) > 0, & \rho_a < \rho \le 1. \end{cases}$$

Consequently every global maximizer belongs to

$$\left\{\rho^*,\ 1,\ 1 - \sqrt{\frac{Cd}{\mu q_1(v_1 - v_2)}}\right\},$$

where the third candidate is retained only if it lies in (ρ^*, ρ_a). In particular, ρ^* is the unique optimum if

$$\mu q_1(v_1 - v_2) < \frac{Cd}{(1-\rho^*)^2} \quad \textit{and} \quad W(\rho^*) > W(1).$$

If $\rho^* > 1$, welfare is strictly increasing throughout the feasible interval.

Up to the threshold, both tiers fill in stage 1 and every surviving vacancy is used: welfare rises linearly even along the exhaustion locus. Past the threshold, the first term in the middle derivative is the gain from replacing delayed safety assignments with delayed reach assignments as more reach vacancies survive; the second is the delay cost magnified by the endogenous expansion of reaching, the moral-hazard cost of fuller insurance. The middle derivative is strictly decreasing in ρ. Once the full-insurance root is reached, the volume rescued is fixed and further reliability improves only the tier composition of rescues, so welfare rises again. The behavioral threshold and the welfare optimum therefore coincide under an economically interpretable condition on primitives.

The constant $C = q_1 + q_2 - (1-\mu)$ that scales the delay penalty is the capacity-ampleness margin: the threshold is interior exactly when $C > 0$. The welfare penalty beyond the threshold therefore arises precisely in the markets whose threshold is interior.

The next corollary turns this finite candidate set into a selection rule. It also distinguishes the point at which two candidates merge algebraically from the parameter value at which the optimal policy changes.

Corollary 2 (When to cap rescue: exact welfare selection under score priority)**.** *Suppose $\rho^* < 1$ and maintain Assumption 3. Let*

$$\Delta = v_1 - v_2, \qquad A = \mu q_1 \Delta, \qquad x = 1-\rho^*, \qquad u = 1-\rho_a, \qquad y = \sqrt{\frac{Cd}{A}},$$

and write $\hat\rho = 1-y$. Then:

(i) every global optimum belongs to $\{\rho^*, \hat\rho, 1\}$, with $\hat\rho$ retained only when $u < y < x$ (equivalently, $\rho^* < \hat\rho < \rho_a$);

(ii) the two comparisons with complete reliability have the closed forms

$$W(1) - W(\rho^*) = Ax - (1-\mu)d(q_2 - s_a^*),$$

$$W(1) - W(\hat\rho) = Ay\left(2 - \frac{y}{u}\right) \quad \text{when } u < y < x;$$

(iii) away from ties, the optimum is

$$\rho^W = \begin{cases} \hat\rho & \text{if } 2u < y < x \quad \text{(throttle)}, \\ \rho^* & \text{if } y \geq x \text{ and } (1-\mu)d(q_2 - s_a^*) > Ax \quad \text{(absorb)}, \\ 1 & \text{otherwise} \quad \text{(complete)}. \end{cases}$$

If $y = 2u < x$, both $\hat\rho$ and 1 are optimal. If $y \geq x$ and $(1-\mu)d(q_2 - s_a^*) = Ax$, both ρ^* and 1 are optimal;

(iv) the closed-form value

$$d_2 = \frac{Ax^2}{C} = \frac{\mu q_1 (v_1 - v_2)(1-\rho^*)^2}{C}$$

is the algebraic value where $\hat\rho$ reaches ρ^*. Suppose $d_2 < d_4$ and the remaining clauses of Assumption 3 hold in a neighborhood of d_2, so the merger lies inside the characterized domain. It is then an actual throttle–absorb boundary if and only if the capped candidate beats completion there, equivalently $x > 2u(d_2)$. Under this strict inequality, throttle is uniquely optimal just below d_2 and absorb is uniquely optimal just above it. If $x < 2u(d_2)$, complete reliability is strictly optimal at d_2; at equality ρ^* and completion tie. If $d_2 \geq d_4$ or another maintained clause fails at d_2, the algebraic merger lies outside this corollary's domain and carries no throttle–absorb conclusion;

(v) complete reliability is optimal for all sufficiently small d. Moreover, let

$$d_4 = \Delta \frac{p_m(q_2)}{1 - p_m(q_2)}.$$

If the remaining clauses of Assumption 3 continue to hold as $d \uparrow d_4$, complete reliability is also optimal for all d sufficiently close to d_4 from below. Thus capped reliability can arise only at intermediate delay costs.

The optimum never lies in the interior of the oversubscribed exhaustion locus: there every unit of reliability rescues a claimant at positive net value, so the designer never stops partway. Past the threshold she can stop where the marginal delay cost catches the recomposition gain (*throttle*), stop at the threshold itself (*absorb*), or run to full reliability (*complete*). Two dimensionless comparisons decide among them: the ratio y/u determines whether the throttle candidate beats completion, and the delay-loss statistic $(1-\mu)d(q_2 - s_a^*)$ determines whether absorption beats completion. The algebraic point d_2 marks where the two capped candidates merge; it lies inside the characterized domain when $d_2 < d_4$ and the remaining maintained conditions hold, and becomes a policy boundary when the dominance test in part (iv) is satisfied.

Remark 1 (Boundary completion and four-regime selections). The threshold-ordering clause $s_a^* < q_2$ is equivalent to $d < d_4$. As $d \uparrow d_4$, $s_a^* \uparrow q_2$, $\rho_a \downarrow \rho^*$, and $W(1) - W(\rho^*) \to Ax > 0$. Hence complete reliability is optimal near both boundaries of the nondegenerate delay-cost range. If $d_2 < d_4$, the remaining maintained clauses hold around d_2, $x > 2u(d_2)$, and the two comparison equations cross only once, then the resulting path is complete–throttle–absorb–complete, as at the three configurations in Table B6. At $(q_1, q_2, \mu, v_1, v_2, \varepsilon) = (0.30, 0.60, 0.20, 1, 0.80, 0.15)$ every maintained inequality holds and complete reliability is optimal throughout $(0, d_4)$; Online Appendix B reports the calculation (Table B6). At d_4 the full-insurance root reaches the safety-subscription boundary, safety rejects enter the clearing problem, and claimant composition changes. Reach-tier vacancies cover the margin at a separate, parameter-dependent threshold when that threshold is feasible.

Online Appendix B verifies the characterization against direct computation. With melt $\mu = 0.05$ the market sits at the knife edge $\rho^* = 1$: the closed-form exhaustion locus matches the computed equilibrium to four decimal places below one, while full service at the endpoint selects the full-insurance root (Table B1). Raising melt to $\mu = 0.062$ moves $\rho^* = 0.7963$ inside the unit interval. The computed threshold is continuous at ρ^* and follows the two closed-form pinned branches on either side. Welfare equals 0.78755 at ρ^*, compared with 0.78298 at $\rho = 1$, and the right derivative at ρ^* is -0.01198; both sufficient conditions in the corollary hold (Table B2). Table B6 reports the transition roots at three configurations where the four regimes occur and a fourth where complete reliability is optimal throughout.

Theorem 1 holds score priority fixed. Even when reliability cannot be reversed after vacancy information is published, the designer can change which claimants are served before the marginal reach applicant. Section 4 measures the protection this reordering affords.

4 The Limits of Aftermarket Design

Once the main round closes, melt and technology determine the stock of surviving seats. The designer still controls claimant order, delay, residual-seat quality, and the credibility of any cap on utilization. How much discipline can these instruments restore? Theorem 2 characterizes the limits of priority design; Proposition 1 varies the quality of rescue; Theorem 3 endogenizes the claimant buffer; and Proposition 2 studies implementation of a welfare-improving cap.

4.1 The buffer principle and an impossibility

Among the three protocols of Section 2, only buffered priority holds the marginal reach applicant to the back of the entire market. Comparing the protocols requires one restriction on admissible designs.

Definition 2 (Non-wastefulness)**.** A protocol is *non-wasteful* if every surviving seat for which an eligible, willing claimant remains is assigned.

The restriction rules out the one trivially effective design: burning seats. A designer who could commit to discard surviving supply could synthesize any reliability she liked; the commitment analysis later in Section 4 shows that ex-post rationality forbids it. Within the non-wasteful class, we measure a protocol's protective power by the largest reliability at which the cautious outcome survives as an equilibrium.

Theorem 2 (The limits of priority design)**.** *(a) (*Buffer principle*) For each protocol π, the cautious outcome is an equilibrium on an interval $[0, \rho_\pi]$, and*

$$\rho_{\text{lot}} = 0, \qquad \rho_{\text{score}} = \min\{\rho_0, 1\}, \qquad \rho_{\text{buf}} = \min\{\rho^*, 1\}.$$

*Before truncation at one, $0 < \rho_0 < \rho^*$ under Assumption 3. Buffered priority attains the maximal feasible threshold. At a boundary $\rho = \rho_\pi$ the statement concerns survival of the cautious equilibrium, not uniqueness; the marginal-rationing convention can support additional boundary ACE. (b) (*Impossibility*) For $\rho > \rho^*$, under any non-wasteful protocol the* *cautious outcome is not an equilibrium: aftermarket supply exceeds total stage-2 demand, the marginal reach applicant is rescued with certainty, and*

$$g(s_c^*) \;\geq\; \frac{(v_1 - v_2)(v_2 - d)}{v_1} \;>\; 0.$$

*(c) (*Prices, not priorities*) When the marginal applicant is fully rescued into a safety seat, the full-insurance root satisfies $p_m(s_a^*) = d/(v_1 - v_2 + d)$, strictly increasing in d; making the aftermarket less convenient disciplines reaching at the margin, but full restoration of the cautious outcome would require $d \geq v_2$, outside the maintained payoff domain.*

Part (a) explains why rejects-first works. Under buffered priority a surviving seat must pass through the entire buffer before it can reach the strategic margin: the marginal reach applicant is served only when $S \geq D_f + D_r$. Where that condition first binds is governed by an identity worth displaying. At any threshold with both claimant classes nonempty ($q_2 \leq s^* \leq 1 - q_1$),

$$D_f + D_r \;=\; (1-\mu)\big[(1 - s^* - q_1) + (s^* - q_2)\big] \;=\; (1-\mu)(1 - q_1 - q_2).$$

The threshold cancels. Total stage-2 demand is a market constant in this region: raising s^* converts failed reach applicants into safety-tier rejects one for one. Buffered priority's erosion boundary therefore solves $\rho\mu(q_1 + q_2) = (1-\mu)(1 - q_1 - q_2)$, subject to the feasible cap $\rho \leq 1$. Score priority erodes earlier at ρ_0 when that threshold is feasible; a lottery grants the marginal reach applicant positive service probability as soon as any seat survives.

Part (b) is the negative counterpart. Past ρ^*, supply exceeds $D_f + D_r$: seats remain after every reject and every higher-scoring failed applicant has been offered one. Non-wastefulness then forces the rescue of the marginal reach applicant with certainty, and the marginal incentive at the cautious root is strictly positive for every $d < v_2$. The disciplining resource is the stock of claimants who can be interposed between surviving seats and strategic entrants; at ρ^* that stock runs out.

The same logic yields a bound in units of buffer capacity.

Corollary 3 (Buffer-capacity bound)**.** *For a protocol π, define its* buffer capacity *$B(\pi)$ as the largest surviving supply at which the marginal reach applicant is served with probability zero at the cautious root. For the fixed claimant orders studied here, the cautious outcome is* *an equilibrium exactly on* $[0, \rho_\pi]$ *with*

$$\rho_\pi = \min\Big\{\tfrac{B(\pi)}{\mu(q_1+q_2)},\, 1\Big\},$$
$$B(\text{lot}) = 0, \qquad B(\text{score}) = (1-\mu)(1-q_1-s_c^*),$$
$$B(\text{buf}) = (1-\mu)(1-q_1-q_2),$$

and every non-wasteful protocol satisfies $B(\pi) \le (1-\mu)(1-q_1-q_2)$.

Priority design determines which claimants stand between surviving supply and the strategic margin; in other institutions the buffer may be wait-listed incumbents, displaced employees, or students from canceled programs. Once surviving supply exceeds the mass a fixed order can place ahead of the marginal strategic claimant, a non-wasteful aftermarket must begin insuring that margin.

Part (c) turns from ordering to prices. Conditional on full safety-seat insurance of the marginal applicant, a higher delay cost moves the full-insurance root toward caution because reaching then requires a higher admission probability; full restoration requires $d \ge v_2$, at which point a rescued safety seat has no positive value. Delay can reduce reaching at the margin but cannot restore caution while rescue remains valuable. Institutionally this lever is the calendar: a later supplementary round raises d without touching the queue.

4.2 The quality of the threat

Theorem 2(b) treats safety-tier fallback seats as interchangeable, so a rescued applicant loses only the delay d. Residual seats may instead be poor matches for someone who did not choose them—unpopular programs, remote campuses, unfunded positions. To ask whether this harsher threat can restore what ordering cannot, we let a failed reach applicant's value of a residual safety seat vary with its depth, while a deliberate main-round safety application retains the guaranteed value v_H; the ladder measures mismatch, not a common-value ranking.[4] The proposition sets $\rho = 1$ and serves the marginal reach applicant last, the configuration where the homogeneous-seat impossibility is strongest.

Proposition 1 (The quality of the threat)**.** *Take the baseline safety value to be* $v_2 = v_H$*, so a main-round safety application yields* v_H*. Index the residual safety inventory by depth* $x \in [0, q_2]$*, and let a failed reach applicant value a seat at that depth at* $v(x) = v_H - (v_H - v_L)x/q_2$*,*

[4] If the ladder were instead a common-value ranking and main-round safety applicants were assigned assortatively along it, application sets need not be thresholds: sufficiently low-score students can prefer certain failure followed by a better residual seat to their low-ranked main-round seat. Proposition 1 treats residual-seat depth as applicant-specific match quality and retains threshold applications.

where $v_1 > v_H > d > 0$ and $v_H \geq v_L \geq 0$. Stage-1 safety choices occupy depths $[0, s^*]$, and stage 2 offers vacancies from the top of the residual inventory. Set $\rho = 1$. Reach-tier vacancies are assigned before safety-tier vacancies; failed reach applicants are ranked by score, so the marginal failed reach applicant is served last. Let $a = 1 - \mu$,

$$x_0 = a - q_1, \qquad s_R = 1 - \frac{q_1}{a},$$

let s_c^H solve $p_m(s_c^H) = v_H/v_1$, and suppose $0 < x_0 \leq q_2$, $s_c^H < q_2$, and either $d \neq v_1 - v_H$ or $p_m(s_R) > 0$. Whenever $s^* < s_R$, the marginal failed applicant reaches the safety ladder at depth

$$x(s^*) = \begin{cases} x_0, & s^* \leq x_0, \\ \dfrac{a(1 - s^*) - q_1}{\mu}, & x_0 < s^* < s_R. \end{cases}$$

Thus depth is conserved on the deep branch and improves endogenously on the shallow branch. Her continuation value is $W(s^*) = \max\{0, v(x(s^*)) - d\}$, and any interior equilibrium on these branches solves

$$p_m(s^*) = \frac{v_H - W(s^*)}{v_1 - W(s^*)}.$$

The stable equilibrium is unique (allowing a pinned solution at s_R) and, holding v_H fixed, its threshold is weakly increasing as v_L falls. The increase is strict whenever the equilibrium lies below s_R and $W(s^*) > 0$. The cautious outcome is fully restored provided $s_c^H < s_R$ and $v(x(s_c^H)) \leq d$. If $s_c^H \geq s_R$, reach-tier melt vacancies alone insure the marginal applicant no later than the cautious root; at equality, total surviving supply still offers the boundary type a positive-value seat. Hence safety-seat dispersion cannot fully restore caution.

The disjunction in the proposition excludes a neutrality knife edge. If $d = v_1 - v_H$ and reach-vacancy coverage begins while $p_m = 0$, then a delayed reach seat is worth exactly the deliberate safety option and an interval of thresholds can be indifferent. For example, at

$$(q_1, q_2, \mu, \varepsilon, v_1, v_H, v_L, d) = (.40, .90, .40, .25, 1, .80, .80, .20),$$

$s_R = 1/3$, $p_m(s) = 0$ through $s = .35$, and every threshold in $[1/3, .35]$ is neutral under the tie-breaking convention. Away from this equality, the crossing is strict.

The conservation identity counts both sources of rescue. A mass μq_1 of surviving reach-tier vacancies is used before the marginal applicant reaches the safety ladder. Safety vacancies weakly above depth x then add $\mu \min\{x, s^*\} + (x - s^*)^+$. Equating their sum to D_f yields $x_0 = a - q_1$. On the shallow branch a more cautious threshold creates fewer failed applicants, so the marginal applicant receives a better residual seat. This feedback is part of the fixed point—it cannot be replaced by the deep-branch constant—and it dampens dispersion: greater caution improves the rescue awaiting the marginal applicant.

Table B4 traces the fixed point; the homogeneous column reproduces the full-insurance root. Dispersion raises the threshold from 0.6417 to 0.7661 over the reported range. It does not reach the cautious root 0.7900 at these parameters because $s_R = 0.7868 < s_c^H$: before caution is restored, reach-tier melt vacancies alone cover all failed reach applicants—the quality instrument's limit.

Main-round discipline depends on the gap between the seat a safety application secures and the seat a failed reach delivers, so the aftermarket menu is itself an incentive instrument: more attractive residual seats are better insurance and invite reaching, though lowering safety-tier quality cannot restore caution once reach-tier melt vacancies cover the strategic failures. Institutionally the lever is the posted vacancy list: replenishing it with desirable seats blunts the threat.

4.3 Endogenous buffers and hysteresis

Buffered priority earned the maximal threshold in Theorem 2(a), but the buffer itself is an equilibrium object. Rejects exist only when the safety tier is oversubscribed, which depends on how aggressively students reach: the protocol protects discipline with a resource that discipline itself produces.

Theorem 3 (Endogenous buffers and hysteresis)**.** *Consider buffered priority with random order within the class of failed reach applicants (the* scramble variant*: classes are ordered, but service within the last class is a lottery). There is an open set of parameters at which a cautious and an aggressive stable equilibrium coexist, separated by an unstable threshold. The buffer that sustains discipline exists only in the cautious regime: aggressive play leaves the safety tier undersubscribed, eliminates rejections, and thereby dissolves the buffer, making aggressive play self-confirming. Under threshold best-response dynamics, transitory shocks can therefore tip the market permanently.*

In the cautious regime $s_c^* > q_2$, the safety tier is oversubscribed and its rejects keep surviving seats away from the strategic margin. Under aggressive reaching the safety tier falls below capacity, $D_r = 0$, and buffered priority loses this first claimant class. The within-class lottery then gives the marginal reach applicant positive service probability, including access to surviving reach-tier seats. Table B5 in Online Appendix B reports, at $\mu = 0.062$ and $\rho = 0.60$, a cautious equilibrium at $s^* = 0.7900$, an aggressive equilibrium at 0.7434, and an unstable threshold at 0.7790 between them.[5]

[5]The coexistence calculation evaluates the explicit incentive function on both sides of each root. All sign

Remark 2. The within-class ordering is essential. Under score order within the last class, the marginal reach applicant is served after every other failed applicant, and the aggressive root does not exist at the parameters of Table B5. Random order within the class, as in a decentralized scramble, restores partial insurance at the margin and permits self-confirming aggression.

Under the threshold best-response dynamics used in the proof, a transitory surge in reaching can carry the market across the unstable threshold. When the shock recedes, the primitives return but the equilibrium need not, because the reject buffer has dissolved.

4.4 Commitment: rescue as bailout

Reliability is partly a policy choice: posting speed, application windows, and quota-lapse rules affect ρ. Corollary 1 identifies the ex-ante optimum, but its implementation is time-inconsistent: once applications are sunk, every additional rescue raises realized welfare by $v_j - d > 0$. Non-wastefulness in Theorem 2(b) can therefore arise from the designer's own ex-post incentives.

Suppose technology makes the seat stock corresponding to $\bar{\rho}$ available independently of the designer's announcement. Before the main round she announces a utilization cap $\hat{\rho} \leq \bar{\rho}$—a promise to let some otherwise available vacancies lapse—and, absent commitment, retains discretion to exceed that cap after melt is realized. This timing distinguishes an ex-post rescue decision from an irreversible ex-ante investment in vacancy detection.

Proposition 2 (Rescue as bailout)**.** *Fix score priority. Let the designer choose an announced utilization cap $\hat{\rho} \leq \bar{\rho}$ before the main round and retain ex-post discretion over rescues, with payoff equal to realized welfare. (i) Without commitment, in any subgame-perfect equilibrium the designer rescues every eligible claimant with a technologically surviving seat; anticipating this, students play the equilibrium of Theorem 1 at $\bar{\rho}$. (ii) With commitment, the designer chooses $\rho^W(\bar{\rho}) \in \arg\max_{\rho \in [0,\bar{\rho}]} W(\rho)$. Under the two strict sufficient conditions in Corollary 1, $\rho^W(\bar{\rho}) = \min\{\rho^*, \bar{\rho}\}$, and the value of commitment is strictly positive if and only if $\bar{\rho} > \rho^*$.*

Part (i) makes the announced cap cheap talk: students expect all technologically surviving seats to be used and play the equilibrium at $\bar{\rho}$. The mechanism has the bailout logic of Kydland and Prescott (1977) and Farhi and Tirole (2012): ex-post rescue changes ex-ante exposure. Part (ii) separates this time-consistency result from the parameter-dependent welfare optimum. At the benchmark, ρ^* is uniquely optimal and commitment has value when

inequalities are strict, so continuity extends the result to an open parameter neighborhood.

$\bar{\rho} > \rho^*$. When delay is cheap enough for complete reliability to maximize welfare, a rescue cap has no commitment value.

Calendar lapse rules, short application windows, and eligibility restrictions make a utilization cap difficult to reverse after applications are submitted. Under the benchmark welfare inequalities, these rules can implement $\min\{\rho^*, \bar{\rho}\}$, improving welfare when the ex-ante discipline they create exceeds the value of the rescues they prevent.

The analysis so far restricts eligibility to unmatched claimants. Allowing assigned students to trade up turns one vacancy into a chain of vacancies, the setting of Section 5.

5 Open versus Closed Aftermarkets

SOAP and Chinese supplementary admissions are *closed*: participation is restricted to unmatched applicants. American college waitlists are *open*: a student holding a deposit may accept another offer, and the vacated seat moves down the quality ladder. Open eligibility multiplies placements mechanically. This section asks when it creates net value: when sorting gains exceed switching costs and the expected losses from chain mortality and delay.

We isolate the eligibility margin in a ladder module that holds main-round behavior fixed.[6] Seats sit on $R \geq 2$ rungs $k = 1, \ldots, R$ with values $v_1 > \cdots > v_R$ and gaps $\Delta_k = v_k - v_{k+1}$; with $R = 2$ the ladder is the reach–safety structure of Section 2.[7] The main round has filled the ladder assortatively, the incumbent at rung k carrying score s_k, decreasing in k, with an unassigned agent of score $s_u < s_R$ outside. Condition on a top vacancy having survived into the aftermarket. Closed eligibility hands it directly to the unassigned agent at value $v_1 - d$. Open eligibility lets the rung-2 incumbent claim it at an all-in switching cost $c > 0$ (including the cost of a late move), vacating rung 2, and so on down the ladder; the previously unassigned terminal claimant bears delay cost d. Each newly created vacancy survives to the next link with probability ρ. We assume $\Delta_k > c$ and $v_R > d$ in the common-value benchmark.

The chain arithmetic relates observed placements to the initial vacancy.

Proposition 3 (Chain arithmetic)**.** *With an open aftermarket, a surviving unit vacancy at the top of an R-rung ladder generates $(1 - \rho^{R-1})/(1 - \rho)$ expected incumbent upgrades and $(1 - \rho^R)/(1 - \rho)$ expected total placements, including terminal absorption. It dies before absorption with probability $1 - \rho^{R-1}$ and is absorbed by an unassigned agent with probability ρ^{R-1}. At $\rho = 1$, the two expectations are $R - 1$ and R.*

[6] The welfare comparison is per unit vacancy within the aftermarket; the feedback from eligibility rules to main-round reports is the subject of the remark that closes this section.

[7] Within this section v_k denotes the value of rung k.

The amplification factor for total observed placements is $(1-\rho^R)/(1-\rho)$ and rises toward $1/(1-\rho)$ as the ladder lengthens: one melted student can generate several observed moves, so measured turnover in an open aftermarket is a multiple of the attrition that caused it. The proposition also prices fragility: absorption requires all $R-1$ subsequent links to survive, so long chains deliver seats to the unassigned only in highly reliable environments.

To isolate the resource cost of chains, suppose all agents derive the same value from a given seat.

Proposition 4 (Common values: closed dominates)**.** *With common values, per unit vacancy, $W_{\text{closed}} - W_{\text{open}} = (R-1)c$ at $\rho = 1$, and strictly larger for $\rho < 1$: vacancy chains are pure waste absent complementarities.*

The accounting is immediate:

$$W_{\text{open}} = \sum_{k=1}^{R-1} \rho^{k-1}(\Delta_k - c) + \rho^{R-1}(v_R - d), \qquad W_{\text{closed}} = v_1 - d.$$

At $\rho = 1$ the movers' gross gains telescope to $v_1 - v_R$, exactly the value the unassigned agent gives up in moving from the top seat to rung R. The reshuffling is a wash before costs; open eligibility is left holding $R-1$ switching costs. For $\rho < 1$, mortality further worsens the comparison.

The case for open eligibility therefore rests on complementarities: chains create surplus only if the agents climbing the ladder are better matched to high rungs than the agent who otherwise receives the top seat. Multiplicative match values make the comparison exact.

Proposition 5 (The sorting value of chains)**.** *With match values $f(v,s) = v \cdot s$, suppose $s_{k+1}\Delta_k > c$ for $k = 1, \ldots, R-1$ and $s_u v_R > d$. The exact expected welfare difference is*

$$W_{\text{open}} - W_{\text{closed}} = \sum_{k=1}^{R-1} \rho^{k-1}(s_{k+1}\Delta_k - c) + \rho^{R-1}(s_u v_R - d) - (s_u v_1 - d).$$

At $\rho = 1$ this becomes

$$W_{\text{open}} - W_{\text{closed}} = \sum_{k=2}^{R} (s_k - s_u)\Delta_{k-1} - (R-1)c.$$

Thus the gross sorting value is the incumbents' score premium over the unassigned, weighted by rung gaps. Open eligibility is strictly preferred exactly when the displayed expected difference is positive, and is weakly optimal when it is nonnegative.

At $\rho = 1$, each term $(s_k - s_u)\Delta_{k-1}$ pairs the score premium of the rung-k incumbent over the unassigned agent with the quality gap that incumbent climbs. For $\rho < 1$, the survival weights ρ^{k-1} account for chain mortality. Open eligibility is more valuable when score premia are large, rung gaps wide, and switching costs low.

The institutional comparison maps directly into these terms. American undergraduate admissions combine a long summer window with peer and program complementarities, and NACAC's 2019 removal of ethics-code provisions barring recruitment of committed students lowered switching costs. Time-critical systems such as SOAP and Chinese supplementary admissions use windows measured in days and allow quotas to lapse, which raises the mortality and delay terms. Signing the welfare comparison requires measures of score premia, rung gaps, survival, and switching costs.

The ladder module conditions on main-round behavior; endogenizing it introduces a second application distortion.

Remark 3 (Seat parking). Open eligibility grants incumbents a trade-up option: a student may secure a safe seat in the main round and climb the ladder through the aftermarket. This option can induce strategically conservative reports, or *seat parking*. Rescue insures aggressive applications, whereas upgrade options subsidize caution. A model that endogenizes parking would determine its interaction with the exhaustion locus of Theorem 1.

6 Empirical Implications and Institutional Evidence

Clearinghouses publish round-specific cutoffs and aftermarket volumes, procedure manuals record eligibility and priority rules, and cohort and capacity series determine the excess-demand term in ρ^*. This section derives three implications from these observables: cutoff reversal, whose incidence we document in official Jiangsu admissions tables; a dated kink in aftermarket volume; and an aggregate-risk correction to the insurance comparison. It closes by mapping four institutions into the model's design variables.

6.1 Cutoff reversal

Common intuition treats second-chance rounds as markets for weak leftovers, so cutoffs should fall. In a score-priority equilibrium, failed reach applicants carry scores of at least s^*, whereas the main-round safety cutoff is $(s^* - q_2)^+$. Whether this ordering determines the aftermarket cutoff turns on how far supply extends into the lower-scoring safety rejects.

Proposition 6 (Cutoff comparison under score priority)**.** *Suppose a positive mass of failed reach applicants receives safety-tier seats in a score-priority ACE. If $s^* \le q_2$ or $S \le D_f$, the* *aftermarket safety cutoff strictly exceeds the main-round safety cutoff. If instead $s^* > q_2$ and $S > D_f$, a positive mass of safety rejects also receives safety-tier seats, and the aftermarket safety cutoff is strictly below the main-round safety cutoff. Along the equilibrium path of Theorem 1, only the first case occurs.*

Reversal identifies which claimant class sets the aftermarket cutoff. In the oversubscribed exhaustion phase all failed reach applicants are served and no safety reject is served; the two cutoffs are s^* and $s^* - q_2$, so their gap is exactly q_2. Beyond the threshold they are s^* and zero, so the gap equals s^* and falls as reaching expands. Along the characterized path, reversal is present and its magnitude tracks the phase. With sufficiently abundant supply, safety rejects also enter the aftermarket and set a cutoff below the main-round cutoff even while failed reach applicants receive safety seats.

Chinese provinces publish cutoffs for both rounds, so the share of programs displaying *daohua* is directly observable. Using official Jiangsu ordinary-undergraduate-batch tables, we find a strictly higher supplementary-round minimum score in 138 of 326 comparable institution–program groups (42.3 percent) in 2023, 190 of 333 (57.1 percent) in 2024, and 225 of 336 (67.0 percent) in 2025.[8]

6.2 Regime-switch dates

The threshold of Theorem 1 is a market statistic: restoring the cohort mass N turns it into a time series. Demographic decline shrinks N, enrollment expansion raises $q_1 + q_2$, and both compress the excess demand that keeps a market on the safe side of the threshold.

Proposition 7 (Regime-switch dates)**.** *Under the rescaled analogues of Assumptions 2 and 3, with cohort mass N, the threshold is $\rho^*(N) = (1-\mu)(N - q_1 - q_2)/(\mu(q_1 + q_2))$, increasing in excess demand. Under fixed institutional reliability $\hat{\rho}$, shrinking excess demand produces a calculable date at which $\rho^*(N)$ falls below $\hat{\rho}$. If $0 < \hat{\rho} < 1$, the equilibrium threshold and aftermarket volume are continuous at that date, but volume changes from locally constant in N before the crossing to increasing as N falls immediately after it, while the equilibrium remains on the undersubscribed exhaustion branch.*

The proposition converts population projections into dated predictions. Chinese birth cohorts are falling while admission capacity grows, so $\rho^*(N)$ declines year by year at roughly

[8]Author's calculations from the Jiangsu Provincial Education Examinations Authority's filing-score and supplementary-plan tables (Jiangsu Provincial Education Examinations Authority, 2023–2025). The unit is year–subject track–institution code–program-group code; a reversal is a strictly higher supplementary-round minimum. The main sample excludes institutions whose supplementary plans mark fully or partly new capacity. Relative to all regular-round groups, the corresponding incidences are 3.8, 4.8, and 5.2 percent. The samples are repeated cross-sections. The replication package contains the official source links, matching code, row-level matches, and sensitivity samples.

fixed institutional reliability. The model predicts a calculable crossing year, marked by unfilled main-round safety seats and a slope change in supplementary volume; because the crossing is generically a kink, it must be read from slopes rather than levels.[9] When structural shortage exceeds melt, the model places the threshold above one and predicts that even a large increase in ρ leaves main-round behavior unchanged. Volume alone remains an incomplete statistic on the exhaustion locus and in open systems: demand expands to meet supply and chains multiply placements.

6.3 Aggregate risk and the information comparison

The baseline model makes admission risk idiosyncratic. Real cutoffs move with common shocks: application waves, quota changes, the big-year–small-year cycles familiar in Chinese admissions. A reach applicant then faces admission and rescue probabilities that both depend on the aggregate state; their comovement matters for the insurance calculus at the center of Theorem 1.

Aggregate risk as discipline. Let admission and rescue probabilities (p_ω, r_ω) be comonotone across aggregate states, so both are weakly higher in slack states. The marginal insurance term then satisfies

$$\mathbb{E}[(1-p_\omega)r_\omega] = \mathbb{E}[1-p]\,\mathbb{E}[r] - \mathrm{Cov}(p,r) \leq \mathbb{E}[1-p]\,\mathbb{E}[r].$$

Expanding the expectation yields the identity, and comonotonicity gives $\mathrm{Cov}(p,r) \geq 0$. The aftermarket is thinnest exactly when it is needed most: rescue pools idiosyncratic risk but cannot absorb a common shock, so aggregate uncertainty weakens insurance relative to an independent-marginals benchmark. The covariance is the sufficient statistic for this correction. A state-contingent application model would determine how information revealed before applications changes reaching and cutoffs.

6.4 An institutional map

Table 1 maps the four institutions from the introduction into eligibility, priority, and the determinants of reliability.

SOAP's effect on main-round behavior depends on whether residency capacity and melt satisfy the capacity-ampleness inequality. The short windows and lapsing quotas of Chinese supplementary admissions lower effective reliability and can implement the commitment

[9] Implementing the dating exercise requires province-level cohort, quota, melt, and reliability series.

Table 1: Four aftermarkets in the model's coordinates

Market	Eligibility	Priority	Determinants of reliability ρ	Model implication
NRMP/SOAP (US residency)	Closed: unmatched applicants only	Program preference in timed offer rounds	Formalized timed rounds since 2012	Threshold location depends on structural shortage relative to melt
Supplementary round (China)	Closed: unassigned students only	Score priority (π^{score})	Short, late windows; unfilled quotas lapse	Lower excess demand moves the market toward ρ^*
Waitlists (US colleges)	Open: admitted students may hold a deposit and wait	Admissions discretion	Long summer windows; lower switching friction after NACAC 2019	Vacancy-chain comparison in Section 5
DoSV (Germany)	Closed: unplaced applicants in post-procedure clearing	University rankings	Coordinated post-procedure clearing	Threshold location depends on capacity and melt

Notes: Eligibility and priority entries follow published procedures. The last two columns translate institutional features into model primitives; quantitative regime classification uses market-specific estimates of ρ, melt, and capacity.

mechanism of Proposition 2. Long college waitlists and lower switching costs move the comparison in Proposition 5 toward open eligibility; score premia, rung gaps, and switching costs determine its sign. Online Appendix Section A.6 adds a supply-side participation margin: weaker off-platform options allow the platform to sustain a larger distributional constraint.

Policy variation. The Chinese parallel-list reform, staggered across provinces between 2008 and 2012, reduced main-round assignment risk (Chen and Kesten, 2017); the model predicts lower supplementary volume after adoption. SOAP's introduction in 2012 increased formalization; if residency markets fail the capacity-ampleness condition, the model predicts little change in main-round behavior despite the increase in ρ. NACAC's 2019 rule change lowered switching costs in an open aftermarket, raising the expected length of vacancy chains in the ladder module.

7 Conclusion

Centralized assignment ends once when the clearinghouse announces its match and again when vacated, lapsed, and unfilled seats find their eventual owners. The aftermarket insures main-round failure and thereby changes what applicants attempt. Along the exhaustion locus in Theorem 1, every additional surviving seat is matched by an additional failed-reach claimant after melt. Past the formalization threshold ρ^*, the safety tier is undersubscribed in the main round and fills partly through the aftermarket.

Serving safety-tier rejects before failed-reach applicants protects main-round discipline exactly up to ρ^*. Theorem 2 shows that no non-wasteful protocol does better: priority can reorder the queue but cannot shorten it. Beyond the threshold, discipline rests on residual-seat quality and, when welfare favors a cap, on commitment not to rescue.

Under score priority, two closed-form comparisons select among completion, throttling, and absorption. Completion is optimal near both boundaries of the nondegenerate delay-cost range; the same comparisons locate the intermediate parameter regions where a cap dominates—where discipline is worth the insurance forgone. In public admissions data, cutoff reversal identifies failed-reach applicants entering the safety aftermarket, while cohort and capacity changes determine a dated kink in supplementary-round volume.

Aggregate risk enters the insurance comparison through the covariance between admission and rescue; seat parking, through the trade-up option. State-dependent cutoffs would determine how prediction tools affect reaching across aggregate states, while endogenous reports under open eligibility would connect parking to the exhaustion mechanism.

Aftermarket rules therefore belong inside the original design problem: how a market ends shapes how its participants play its beginning.

Data and code availability. The accompanying replication package contains the equilibrium solver, theorem-regression tests, and the full reconstruction of the Jiangsu cutoff-reversal statistics, including official source links, matching code, row-level observations, summary files, and sensitivity checks. The underlying admissions tables are public records of the Jiangsu Provincial Education Examinations Authority and the package archives the exact official files used by the reconstruction script.

Funding and competing interests. This work was supported by the National Natural Science Foundation of China (grant number 72503198). The author declares no conflict of interest.

References

Abdulkadiroğlu, A., and T. Sönmez (2003): "School Choice: A Mechanism Design Approach," *American Economic Review*, 93(3), 729–747.

Afacan, M. O., and E. Cumbul (2025): "Waitlist Engineering in Discrete Object Allocations with Outside Option," *International Journal of Game Theory*, 54(1), 1–22.

Agarwal, N., I. Ashlagi, E. Azevedo, C. R. Featherstone, and Ö. Karaduman (2019): "Market Failure in Kidney Exchange," *American Economic Review*, 109(11), 4026–4070.

Agarwal, N., I. Ashlagi, M. A. Rees, P. Somaini, and D. Waldinger (2021): "Equilibrium Allocations under Alternative Waitlist Designs: Evidence from Deceased Donor Kidneys," *Econometrica*, 89(1), 37–76.

Akbarpour, M., A. Kapor, C. Neilson, W. van Dijk, and S. Zimmerman (2022): "Centralized School Choice with Unequal Outside Options," *Journal of Public Economics*, 210, 104644.

Ali, S. N., and R. I. Shorrer (2025): "Hedging When Applying: Simultaneous Search with Correlation," *American Economic Review*, 115(2), 571–598.

Andersson, T., U. Dur, S. Ertemel, and O. Kesten (2024): "Sequential School Choice with Public and Private Schools," *Social Choice and Welfare*, 63(2), 231–276.

Arnosti, N. (2015): "Short Lists in Centralized Clearinghouses," in *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, p. 751.

Arnosti, N., and P. Shi (2020): "Design of Lotteries and Wait-Lists for Affordable Housing Allocation," *Management Science*, 66(6), 2291–2307.

Ashlagi, I., F. Monachou, and A. Nikzad (2025): "Optimal Allocation via Waitlists: Simplicity through Information Design," *Review of Economic Studies*, 92(1), 40–68.

Ashlagi, I., and A. E. Roth (2014): "Free Riding and Participation in Large Scale, Multi-Hospital Kidney Exchange," *Theoretical Economics*, 9(3), 817–863.

Avery, C., and J. Levin (2010): "Early Admissions at Selective Colleges," *American Economic Review*, 100(5), 2125–2156.

Avery, C., and P. A. Pathak (2021): "The Distributional Consequences of Public School Choice," *American Economic Review*, 111(1), 129–152.

Azevedo, E. M., and E. Budish (2019): "Strategy-proofness in the Large," *Review of Economic Studies*, 86(1), 81–116.

Azevedo, E. M., and J. D. Leshno (2016): "A Supply and Demand Framework for Two-Sided Matching Markets," *Journal of Political Economy*, 124(5), 1235–1268.

Blum, Y., A. E. Roth, and U. G. Rothblum (1997): "Vacancy Chains and Equilibration in Senior-Level Labor Markets," *Journal of Economic Theory*, 76(2), 362–411.

Çelebi, O., and J. P. Flynn (2022): "Priority Design in Centralized Matching Markets," *Review of Economic Studies*, 89(3), 1245–1277.

Chade, H., and L. Smith (2006): "Simultaneous Search," *Econometrica*, 74(5), 1293–1307.

Che, Y.-K., and Y. Koh (2016): "Decentralized College Admissions," *Journal of Political Economy*, 124(5), 1295–1338.

Che, Y.-K., and O. Tercieux (2026): "Optimal Queue Design," *Journal of Political Economy*, published online April 2. https://doi.org/10.1086/741622.

Chen, Y., and O. Kesten (2017): "Chinese College Admissions and School Choice Reforms: A Theoretical Analysis," *Journal of Political Economy*, 125(1), 99–139.

De Groote, O., A. Fabre, M. Luflade, and A. Maurel (2025): "Sequential College Admission Mechanisms and Off-Platform Options," NBER Working Paper No. 34105.

Doğan, B., and M. B. Yenmez (2019): "Unified versus Divided Enrollment in School Choice: Improving Student Welfare in Chicago," *Games and Economic Behavior*, 118, 366–373.

Doğan, B., and M. B. Yenmez (2023): "When Does an Additional Stage Improve Welfare in Centralized Assignment?" *Economic Theory*, 76(4), 1145–1173.

Doval, L. (2022): "Dynamically Stable Matching," *Theoretical Economics*, 17(2), 687–724.

Dur, U., and O. Kesten (2019): "Sequential versus Simultaneous Assignment Systems and Two Applications," *Economic Theory*, 68(2), 251–283.

Ekmekci, M., and M. B. Yenmez (2019): "Common Enrollment in School Choice," *Theoretical Economics*, 14(4), 1237–1270.

Farhi, E., and J. Tirole (2012): "Collective Moral Hazard, Maturity Mismatch, and Systemic Bailouts," *American Economic Review*, 102(1), 60–93.

Feigenbaum, I., Y. Kanoria, I. Lo, and J. Sethuraman (2020): "Dynamic Matching in School Choice: Efficient Seat Reassignment After Late Cancellations," *Management Science*, 66(11), 5341–5361.

Gale, D., and L. S. Shapley (1962): "College Admissions and the Stability of Marriage," *American Mathematical Monthly*, 69(1), 9–15.

Garratt, R., and T. Tröger (2006): "Speculation in Standard Auctions with Resale," *Econometrica*, 74(3), 753–769.

Gong, B., and Y. Liang (2025): "A Dynamic Matching Mechanism for College Admissions: Theory and Experiment," *Management Science*, 71(5), 4396–4412.

Grenet, J., Y. He, and D. Kübler (2022): "Preference Discovery in University Admissions: The Case for Dynamic Multioffer Mechanisms," *Journal of Political Economy*, 130(6), 1427–1476.

Haeringer, G., and V. Iehlé (2021): "Gradual College Admission," *Journal of Economic Theory*, 198, 105378.

Haile, P. A. (2003): "Auctions with Private Uncertainty and Resale Opportunities," *Journal of Economic Theory*, 108(1), 72–110.

Huitfeldt, I., V. Marone, and D. C. Waldinger (2024): "Designing Dynamic Reassignment Mechanisms: Evidence from GP Allocation," NBER Working Paper No. 32458. https://doi.org/10.3386/w32458.

Jiangsu Provincial Education Examinations Authority (2023–2025): "Ordinary Undergraduate Admissions: Initial-Round Filing Scores, Supplementary Plans, and Supplementary-Round Filing Scores," official data files. https://www.jseea.cn/.

Kapor, A., M. Karnani, and C. Neilson (2024): "Aftermarket Frictions and the Cost of Off-Platform Options in Centralized Assignment Mechanisms," *Journal of Political Economy*, 132(7), 2346–2395.

Kojima, F., and P. A. Pathak (2009): "Incentives and Stability in Large Two-Sided Matching Markets," *American Economic Review*, 99(3), 608–627.

Kuno, K. (2023): "Strategic Waiting in Centralized Matching: Daycare Assignment," arXiv Working Paper No. 2311.07920, revised January 2025.

Kydland, F. E., and E. C. Prescott (1977): "Rules Rather than Discretion: The Inconsistency of Optimal Plans," *Journal of Political Economy*, 85(3), 473–491.

Leshno, J. D. (2022): "Dynamic Matching in Overloaded Waiting Lists," *American Economic Review*, 112(12), 3876–3910.

Manjunath, V., and B. Turhan (2016): "Two School Systems, One District: What to Do When a Unified Admissions Process Is Impossible," *Games and Economic Behavior*, 95, 25–40.

Narita, Y. (2018): "Match or Mismatch? Learning and Inertia in School Choice," SSRN Working Paper No. 3198417.

Niederle, M., and A. E. Roth (2003): "Unraveling Reduces Mobility in a Labor Market: Gastroenterology with and without a Centralized Match," *Journal of Political Economy*, 111(6), 1342–1352.

Nikzad, A., and P. Strack (2024): "Equity and Efficiency in Dynamic Matching: Extreme Waitlist Policies," *Management Science*, 70(8), 5187–5207.

Roth, A. E. (1984): "The Evolution of the Labor Market for Medical Interns and Residents: A Case Study in Game Theory," *Journal of Political Economy*, 92(6), 991–1016.

Roth, A. E., and E. Peranson (1999): "The Redesign of the Matching Market for American Physicians: Some Engineering Aspects of Economic Design," *American Economic Review*, 89(4), 748–780.

Roth, A. E., and X. Xing (1994): "Jumping the Gun: Imperfections and Institutions Related to the Timing of Market Transactions," *American Economic Review*, 84(4), 992–1044.

Schummer, J. (2021): "Influencing Waiting Lists," *Journal of Economic Theory*, 195, 105263.

Vohra, A. (2025): "Unraveling and Inefficient Matching," *American Economic Journal: Microeconomics*, 17(3), 415–453.

Zheng, C. Z. (2002): "Optimal Auction with Resale," *Econometrica*, 70(6), 2197–2224.