

Implementation and Validation of DA algorithm for Joint Seat Allocation 2015

Summary and Recommendation for 2016

**John Augustine, Surender Baswana, Kalpesh Kapoor, D Durga Prasad,
Meghana Nasre, Utkarsh Patange, Vijaya Saradhi, Neelesh Upadhye**

31 July 2015

This year, the Deferred Acceptance (DA) algorithm was used for the joint seat allocation in IITs, NITs, several IIITs, and other government funded technical institutes of India. NIC and IITK implemented the DA algorithm independent of each other. These implementations were validated by teams from IIT Guwahati, IIT Kharagpur and IIT Madras during the 4 rounds of joint seat allocation from 1st July to 22nd July, 2015. In this document, a summary of the implementation and validation of these implementations is provided along with the recommendations for the year 2016.

Validation Modules

Input sanity tests

Following is the list of input Sanity Tests that were implemented:

Candidate Specific Checks:

1. Each Candidate has a valid Category tag (one amongst GEN, SC, ST, OBC-NCL).
2. Each Candidate has at least one non-zero/ non NULL Rank [JEE (Main) 2015 Paper-1, JEE (Main) 2015 Paper-2 / JEE (Advanced) 2015 (regular / preparatory)].
3. Each Candidate has his/her ranks consistent with his/her category tag (eg: an OBC candidate can have GEN and OBC ranks, but not SC or ST rank).
4. For each Candidate the mandatory fields have only allowed values. Eg: PWD_Status can be on 1 or 2. Similarly, in round 1, all decisions have to be Float (FL), but round 2 onwards, they can be one of Freeze, Float, Slide, Reject. Such checks were made for the following fields: State-Code, Gender, PwD-Status, Nationality, is-Preparatory, ADV-DS, AAT-Status, Color-Blind-Status, One-Eyed-Vision, ADV-Top-20, cat-Change, Marks-ENG-Change, Marks-ARC-Change.

Some of these checks did not pass, but it was ensured for each of them did not pass for valid reasons whenever the checks did not pass. For example, to begin with, marks were not available and it was because CBSE provided marks in 2nd round only. Similarly, there were 81,781 candidates who did not have ANY rank (not-even prep ranks). All such candidates had all their Remark_Symbols to be N which means they were not to be considered for allotment.

Seat Matrix Specific Checks: Seat matrix being meaningful, that is for every program GEN, OBC, SC, ST should be > 0. This check also did not pass for all programs. But all such cases were reported to the respective Institutes and it was made sure that the Institutes agreed with the seat-distribution for those programs.

Candidate specific validation modules

1. Fairness check: If a CatChange \neq 1 candidate does not get a seat in virtual program P in spite of the fact that he/she applied to the program and was eligible for the same, then the closing rank of P must be **strictly** less than the rank of the candidate. In addition, if a candidate X (whatever his/her CatChange value may be) applied to a virtual program P, and he/she was denied a seat for P, then P cannot have dereserved seats or vacancies.
2. Restricted fairness for CatChange=1 candidates: This involves closing rank checks/ min-cut off checks. If a CatChange=1 candidate X applied to a virtual program P, and he/she was denied a seat in P, then P must
 - i. have no CatChange=1 candidate with a worse rank than that of X and
 - ii. have min-cut-off strictly better than the rank of X
3. For rounds 2 and beyond, Decisions of the candidates -- Freeze, Float, Reject are respected.
4. Candidates allotted DS or foreign seats belong to CRL.
5. Candidate allotted in a particular category has valid rank in that category.
6. Allotment category is according to Business Rules 18 and 19.
7. HS/OS quota rule is respected for non-IIT candidates.
8. If allotment is supernumerary (EQ: equality of ranks, MC: min-cutoff, FR: foreign), then the reason for supernumerary has to be justified accordingly.
9. Candidates getting seats in preparatory course have to belong to the preparatory category.

The following table states the respective modules developed and used by the validation teams from IITG, IITKGP, and IITM.

Module no.	IITG	IITKGP	IITM
1	Yes	Yes	Yes
2	Yes	No	Partially Yes (2.ii done)
3	Yes	Yes*	Yes
4	Yes	Yes*	Yes
5	Yes	Yes*	Yes
6	Yes	No	Yes
7	Yes	Yes	Yes
8	Yes (partially)	No	Yes
9	Yes	Yes*	Yes

* These checks were done using SQL Commands at the database level

Program specific validation modules

1. Total candidates allotted to a program (as computed by counting from allotment file) matches the total allotted given in Program statistics file.
2. Total capacity of program given in Program-stats file matches the total capacity of the program given in the seat-matrix file.
3. Capacity of the program after dereservation is maintained. That is, $\text{finalCapacity} = \text{initial-capacity} - \text{dereservFrom} + \text{dereservTo}$.
4. A virtual program where dereservation happened cannot have supernumerary seats.
5. $\text{Total allotted} = \text{finalCapacity} + \text{supernumerary}$.
6. $\text{Total supernumerary allotments (counted from the allotments file)} = \text{total supernumerary in Program-Statistics file}$.
7. The final vacancy has to be only in GEN, SC, ST. This is because of dereservation.

Module no.	IITG	IITKGP	IITM
1	Yes	No	Yes
2	Yes	No	Yes
3	No	No	Yes
4	No	No	Yes
5	No	No	Yes
6	No	No	Yes
7	No	No	Yes

Matching allotment module

The allotments produced by the two softwares were matched for program, rank, category, and all flags. Except for supernumerary reason mismatch (see reason following) every other field of the two allotment was matched. Case for mismatch in supernumerary reason: for example, if in a program there are 2 candidates at closing rank then any one of them can be given supernumerary and marked with EQ; therefore there can be a mismatch between supernumerary reason if checked candidate wise. In all such cases when there was a mismatch in supernumerary reason, it was manually verified that the mismatch is indeed due to the above reason.

Summary of the validation tests conducted from 1st July to 22nd July

The validation was conducted on live data a total of 22 times during the four rounds of the joint seat allocation. These also include the Mock allocation and the final allocations for the 4 rounds. The additional 17 tests that were conducted were very helpful in resolving all the issues with the two softwares. Out of these 8 were conducted before the 1st round itself.

Time taken by the two softwares

Till the end of 3rd round, the IITK software used to take around 25 minutes for computing the allocation whereas the NIC software used to take around 3-4 hours. For the final round, NIC software had “in-memory” implementation. However, it was still twice as slow compared to the IITK software. The reason for this slowness probably was that NIC was not able to use the full power of in-memory implementation due to this last minute optimization.

An important parameter that governs the time taken by the allocation software is the number of runs it performs. The software performs a run whenever there is a virtual program in which de-reservation is possible. The number of runs of the

software was around 5 for all the executions before round 3. However, for the 4th round, the number of runs exceeded and reached 28. NIC had now “in-memory” execution and it took around 8 hours whereas IITK software took 4 hours.

Various issues that came up during validation

- On the first day (1st July), the data from NIC had some fields with NULL values (like marks, Adv_Top_20 field), board names with comma. So Neelesh had to write a query to make the data compatible with the desired format for the IITK implementation. There was a minor change in the seat matrix as well that NIC informed us later.
- There was a mismatch in one of the tests done during the reporting of 1st round. Firstly, the decisions of Foreign candidates were not respected by the NIC allotment. Also the slide from 228 to 232 and vice versa was not allowed by the NIC software (228 is the normal institute code for NITW and 232 is the institute code used to denote NITW supernumerary seats for AP candidates). These issues were fixed by NIC within one day.
- Preparatory course candidates were given ranks from their extended merit list by NIC software. This issue was also fixed by NIC.
- While outputting the ranks, the precision of the rank was lost for ranks having more than one digit after the decimal point. This issue was resolved by using higher precision of numbers for ranks but it led to an increase in the running time of both the softwares.
- The validity of the choices was getting reset to NULL by the NIC software in each round. It did not matter in the 1st round. It was later fixed by NIC.
- There was an MC supernumerary reason mismatch in one of the synthetic data sets. However, it was never exposed in any of the live sets. We must ensure that it is fixed. The issue seemed to be an apparent incorrect supernumerary reason in the NIC allotment.

Recommendations for future implementation and validation

The IIT that will organize JEE (Advanced) for the year 2016 must appoint a team of highly competent persons who will be taking the responsibility of validation and parallel implementation of the DA algorithm. The team is advised to enhance the validation modules that were developed and used for the year 2015. **If IITK implementation has to be used for the year 2016, there must be minimal changes in the business rules of JoSAA for 2016.** In case some business rules have to be changed, the technical committee of the DA algorithm must be consulted for the feasibility of their incorporation in the IITK and NIC implementations. The team must develop a thorough understanding of the

software and the way validation has to be carried out. They must also be able to spot the cause of a problem that may come up during validation process. The activities related to the implementation and validation of the software should start as soon as possible for the year 2016 and NIC should be kept informed about all the activities and plan. In short, the task of parallel implementation and validation is a major task and needs to be taken up promptly and very seriously.

The way validation needs to be carried out in future years

- Validation should be an ongoing process with software development. The idea of keeping software development teams and validation teams independent is good, but the validation of allotment software(s) should possibly begin as soon as the allotment software reaches a reasonably stable state.
- A version control for allotment softwares and validation codes/scripts like the one done by IITK must be followed. There are many free services like bitbucket that can be availed. Whenever a new version is prepared, it must be checked on all the previous test cases. This task must be carried out before releasing the new version.
- Raising of bugs and their resolution should be tracked by someone [say by testing teams that raise it or by development teams].
- Validation programs / modules must include solid input sanity checks. Only preliminary checks were done this time, but they helped to quite an extent initially to spot issues.
- In addition, redundant checks must be built in; for example, we had a seemingly simple check like total capacity of a program in the program statistics file should be equal to total capacity of program in seat-matrix. This check, although mostly redundant, was helpful initially to spot a minor change in seat-matrix (due to which the allotments were initially not matching)

The importance of internal mock allocations

By internal mock allocation, we mean allocations which are never displayed to candidates but done for testing purpose on live data. The idea of having (internal) mock allocation starting on 30th June as soon as preliminary data was available was an excellent one. There were 17 such internal mock seat allocations carried out in the month of July 2015. These internal mock seat allocations helped in fixing various issues related to validation of the two softwares which otherwise would have been extremely difficult. Although going forward, it may be expected

that there will be far fewer issues, still planning on such (internal) mock allocation would be a good idea. These internal mock allocations should be carried out in future as well.

The importance of physical presence at NIC during the validation

For the year 2015, five faculty members (from IITG, IITK, IITM) and one senior technical staff member from IITKGP were present at NIC for most of the time during seat allocation. This was in addition to Utkarsh who was present at NIC throughout the 22-day period. It proved to be very useful. The entire process of allotment and validation is a time bound activity. If there are any issues that arise in either allotment or validation they need to be discussed and resolved at the earliest. Discussion happens seamlessly in the physical presence of teams. However, after Round 1, the physical presence could be minimized -- but we don't think it should be completely avoided. There was crucial dependency on the NIC team to give access to data as soon as it was available. Also, initially the presence of teams together at NIC got us to know everyone and also the expertise of different team members in different tasks required during validation. The last minute grouping of the team and getting to know the team members can be avoided if software development and validation teams work together from initial stages.

For the joint seat allocation for the year 2015, Utkarsh being present at the NIC office all the time did help to a great extent for testing for round 2, round 3, and round 4. Among other things, it helped ensure that the data was transferred promptly from NIC to other teams. If the following things are ensured, then travel and physically being present at the NIC/IIT where the "choice filling-in servers" are located can be minimized greatly:

- Team 1 [which is responsible for collecting choices from the candidates] is willing to hand over data to teams located remotely and there is a protocol mutually agreed upon. Sharing the data on Google drive or Dropbox is not recommended.
- For the week before the the first round of allotment, it will be very useful for all the validation teams to assemble at one place. All team members may work closely together and became comfortable with each other so that during the final allotments it will possible to communicate effectively from their respective institutes.

A parallel implementation of the seat allocation algorithm

A natural question that can arise in one's mind is the following:

Should there be a parallel implementation of the allocation algorithm for future?

The answer to this question by the entire validation and implementation team for the year 2015 as well as NIC is a **Yes**. There must be two softwares developed independently for the joint seat allocation. One implementation should be by NIC and the other implementation should be by IITs. This requirement of two softwares is a must throughout the life of JEE. A match between the two implementations is a big confidence booster. In addition there must be at least 2 teams validating the outputs.

Commendation for NIC

NIC did an excellent job. Their team was very cooperative and supportive for the entire period of joint seat allocation. The implementation team at NIC consisting of Vikas Pandey, Neetu Kumar Verma, and Mangal Singh did an excellent work of implementing the DA algorithm. They addressed all the minor issues that arose during the validation very promptly. The rich experience of Anwar ji and Pratik ji for counselling was very helpful. It would have been very helpful if they had also considered the following suggestions.

- Prior to testing the two softwares on live data, the softwares were tested on synthetic data sets which were generated using a Synthetic Data Generation program. The main issue while testing during this phase was that the NIC program has very strict requirements on the input (especially on the various REM_SYMBOLS). It therefore took several iterations for us to get the generation software create data sets that were "valid" for the NIC software. In addition, the GUI based interface of the NIC software made it cumbersome for testing. Going forward, if a command line interface to the NIC software were to be available, it will greatly help in automated testing.
- The NIC software was taking a huge amount of time. Most of the time of the validation team was wasted in just waiting for the output of the NIC software. It will have been helpful if NIC could improve the running time further.

Dependence on NIC for 2016

It was a very good team work with NIC during the joint seat allocation of 2015. We benefitted immensely from their rich experience of handling counselling for so many years. NIC handled the job of data collection (candidate data, choice filling, reporting center activities). The seat allocation was handled jointly by the IIT Kanpur team and the NIC team. Though NIC did a good job as far as the allocation software is concerned, there was some problem during the choice filling. For many hours during the days of choice filling, their servers were not able to handle the load for choice filling. This also led to extending the deadline of choice filling by one day. Though we requested them to fix this problem, they could not do it satisfactorily. It appears to be due to some lack of competence on their part. Considering all aspects of joint seat allocation for the year 2015, we would like to suggest the following options for the year 2016.

1. We follow the same protocol as we did for the year 2015, that is, data collection by NIC, the seat allocation independently by NIC as well as IITs, and validation solely by IITs. NIC has developed expertise in handling common counselling for many years. There was a very good coordination between the NIC team and the IIT teams this year. So in our view, this is the best option for 2016. However, we must try to ensure that the server overload problem at NIC is fixed. We believe that IITs has the technical expertise to help NIC fix this problem.
2. IITs take over the entire task of joint seat allocation. However, this is a huge task. Taking over entire process by IITs needs dedicated resources, highly committed teams of large size, planning, and time (for testing of such a system). As the entire process of allotment depends on data collection, this option should be exercised only after weighting all its advantages and disadvantages.
3. NIC handles the data collection job, and the seat allocation and validation is done solely by IITs using softwares developed independently. There is a potential risk in going with this option. Since the data collection is done by NIC, and IITs just use the data, what if some error crops up in the data due to some lack of communication? Asking NIC for the allocation software also make them more accountable for the integrity of the data. This year, working as a team for the task of allocation helped in resolving some minor issues. If NIC is not involved in the seat allocation task, this crucial cooperation will be missing.

Whichever option is selected for the year 2016, it is strongly recommended that a large group of faculty members and technical staff from IITs should be constituted for the joint seat allocation for 2016. This year, around 6 faculty members, one outstanding student, and one senior technical staff member put extraordinary efforts to ensure that design & implementation of the DA algorithm, and the validation of softwares is done in the best possible way without leaving any room for error. It is time that IIT system should develop a robust and sustainable solution for this mega task instead of putting extraordinary load on just a few committed people.

Recommendation for CBSE

- The delay in declaring ranks by CBSE was the source of the delay in joint seat allocation. Given the track record of CBSE, it might not be reasonable to expect them to respect the deadline in 2016. In future CBSE should plan well in advance so as to declare the ranks list on time. In case CBSE is not able to meet the deadline for declaring the ranks, provision should be made to allow candidates to fill in choices even without their mains ranks. When the ranks are announced, their invalid choices can be removed. For example, suppose a candidate does not have a rank in ARC but he/she opts for architecture program in NIT, then this choice must be made invalid after the ranks are declared. Note that this may greatly increase the total number of choices filled and NIC should be able to handle such a large database.
- If possible, use of floating points for ranks with more than one digit after decimal point should be avoided.