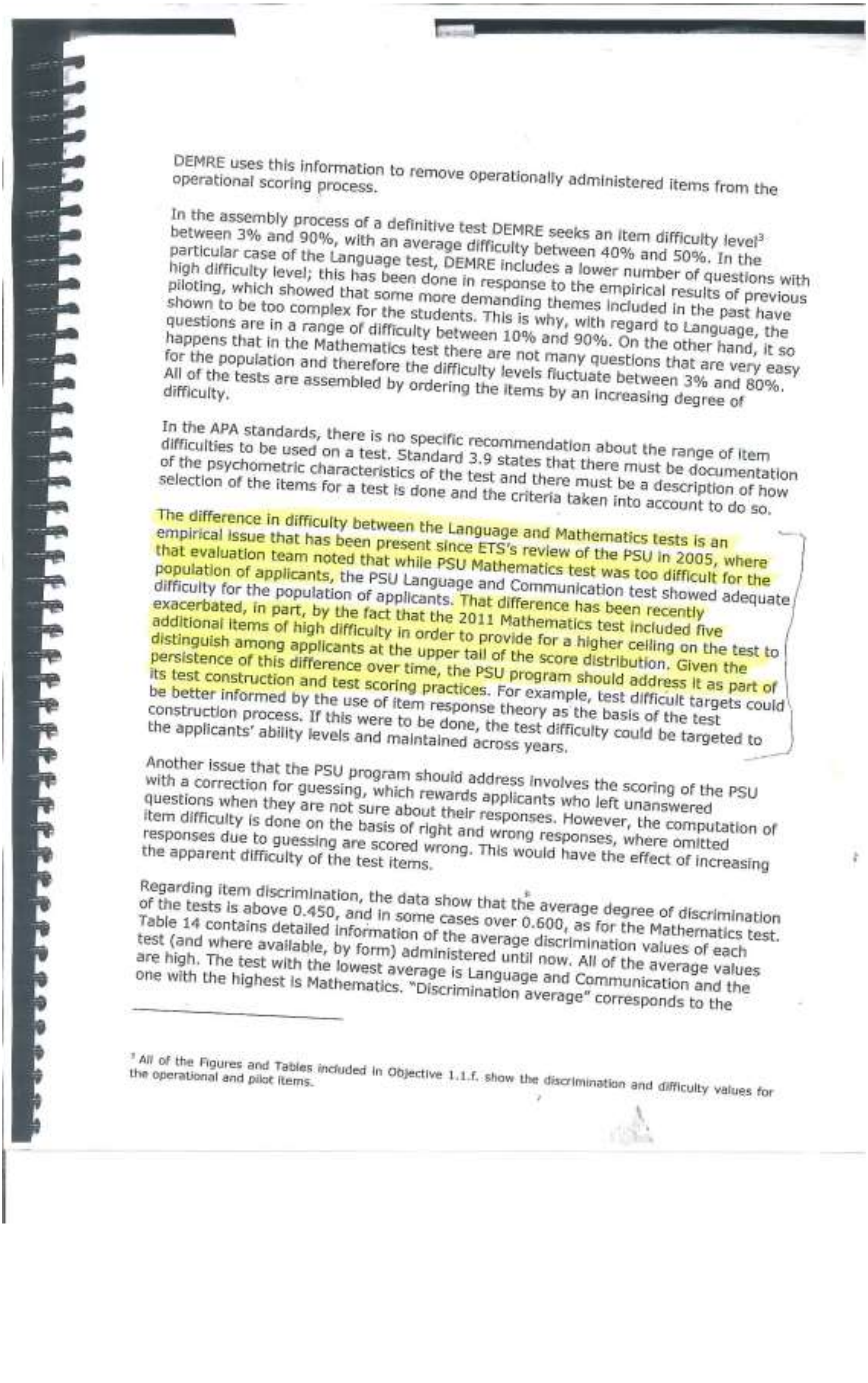
El problema de la PSU de Matemáticas y su agravamiento por malas decisiones del CTA, descrito en el informe de Pearson



DEMRE uses this information to remove operationally administered items from the operational scoring process.

In the assembly process of a definitive test DEMRE seeks an item difficulty level[3] between 3% and 90%, with an average difficulty between 40% and 50%. In the particular case of the Language test, DEMRE includes a lower number of questions with high difficulty level; this has been done in response to the empirical results of previous piloting, which showed that some more demanding themes included in the past have shown to be too complex for the students. This is why, with regard to Language, the questions are in a range of difficulty between 10% and 90%. On the other hand, it so happens that in the Mathematics test there are not many questions that are very easy for the population and therefore the difficulty levels fluctuate between 3% and 80%. All of the tests are assembled by ordering the items by an increasing degree of difficulty.

In the APA standards, there is no specific recommendation about the range of item difficulties to be used on a test. Standard 3.9 states that there must be documentation of the psychometric characteristics of the test and there must be a description of how selection of the items for a test is done and the criteria taken into account to do so.

The difference in difficulty between the Language and Mathematics tests is an empirical issue that has been present since ETS's review of the PSU in 2005, where that evaluation team noted that while PSU Mathematics test was too difficult for the population of applicants, the PSU Language and Communication test showed adequate difficulty for the population of applicants. That difference has been recently exacerbated, in part, by the fact that the 2011 Mathematics test included five additional items of high difficulty in order to provide for a higher ceiling on the test to distinguish among applicants at the upper tail of the score distribution. Given the persistence of this difference over time, the PSU program should address it as part of its test construction and test scoring practices. For example, test difficult targets could be better informed by the use of item response theory as the basis of the test construction process. If this were to be done, the test difficulty could be targeted to the applicants' ability levels and maintained across years.

Another issue that the PSU program should address involves the scoring of the PSU with a correction for guessing, which rewards applicants who left unanswered questions when they are not sure about their responses. However, the computation of item difficulty is done on the basis of right and wrong responses, where omitted responses due to guessing are scored wrong. This would have the effect of increasing the apparent difficulty of the test items.

Regarding item discrimination, the data show that the average degree of discrimination of the tests is above 0.450, and in some cases over 0.600, as for the Mathematics test. Table 14 contains detailed information of the average discrimination values of each test (and where available, by form) administered until now. All of the average values are high. The test with the lowest average is Language and Communication and the one with the highest is Mathematics. "Discrimination average" corresponds to the

[3] All of the Figures and Tables included in Objective 1.1.f. show the discrimination and difficulty values for the operational and pilot items.