Badania

Przygotowanie badań

Wstęp

Po wstępie teoretycznym i części implementacyjnej naturalnym następcą jest część badawcza. Wartość stworzonego środowiska testowego można sprawdzić tylko w jeden sposób – przeprowadzając badania. Część badawczą postanowiono rozpocząć od eksperymentu mającego na celu zadanie sprawdzenie poprawnego działania odtwarzacza, a więc weryfikacje czy wyświetlane wideo jest zgodne z tym co zostało wysłane przez program do karty graficznej. Właściwe testy subiektywne miały na celu porównanie standardowych metod badawczych pod kątem wpływu wyboru metody testu na jego wynik.

Przygotowanie pomieszczenia do badań

Rozpoczynając subiektywne testy oceny jakości wideo należy przygotować odpowiednie warunki do przeprowadzania testów. Standardowe warunki testów zostały zdefiniowane w normach ITU-T Rec. P910 oraz Rec. P913, a także ITU-R Rec. BT.500, co zostało opisane we wstępie teoretycznym. Przygotowując pomieszczenie do przeprowadzania testów należy zwrócić uwagę na wiele aspektów, oczywiście rekomendacje mają charakter zaleceń, a projektujący testy mogą zmieniać założenia w taki sposób, aby wypełniały zadany cel.

Ze względu na fakt braku dostępnego w dowolnych godzinach pomieszczenia do przeprowadzania testów wideo na uczelni, a także pracy zawodowej twórców badań zdecydowano się na organizacje pomieszczenia do testów we własnym zakresie. Odnosząc się do rekomendacji, biorąc pod uwagę konieczny sprzęt oraz możliwości jego przemieszczania określono dwie potencjalne lokalizacje przeprowadzania testów. Rozważono argumenty za i przeciw każdej lokalizacji. Były nimi:

* Sala prób zespołu muzycznego jednego z twórców pracy
* Dom rodzinny drugiego z twórców

Dom ze względu na fakt, że jest miejscem zamieszkania na co dzień jest wyposażony w sprzęt potrzebny do przeprowadzania testów – zarówno komputer posiadający dyski SSD jaki i kartę graficzną, a także telewizor o dobrej jakości obrazu, a także krzesła, fotele i inne meble. W domu jednak zarówno ściany jak i podłoga posiadają nieneutralne kolory, okna mimo zasłon przepuszczają bardzo dużo światła, a samo pomieszczenie posiada różnego rodzaju detale i dekoracje. Dom znajduję się również w sporej odległości od centrum miasta co stwarza dodatkowy problem z organizacją testerów.

Sala jest pomieszczeniem specjalistycznym przygotowanym do tworzenia i nagrywania muzyki. Z tego względu konieczne jest jej odpowiednie wygłuszenie. Ściany zostały pokryte dźwiękoszczelną wełną mineralną. Ze względu na komfort użytkowników, a także sąsiadów okna zostały zasłonięte zarówno wełną jak i twardymi zasłonami nieprzepuszczającymi światła. Całość sali jest utrzymana w jasnych stonowanych szarościach. Pomieszczenie wymaga jednak przygotowania, tymczasowego usunięcia sprzętu muzycznego tak aby nie rozpraszał testerów, a także dostarczenia na miejsce zarówno komputera jak i telewizora. Dodatkową zaletą był bliskość centrum miasta.

Zdecydowano się na sale prób ze względu na większą dostępność, brak konieczności oglądania się na domowników, a także fakt wygłuszenia i neutralnego sztucznego oświetlenia. Na miejsce dostarczono sprzęt. Ponieważ rekomendacja P.910 zakłada użycie dowolnego urządzenia spełniającego założenia badań zdecydowano się na znajdujący się w posiadania jednego z twórców telewizor Samsung wspierający FullHD. Za urządzenie bazowe do uruchomienia oprogramowania posłużył komputer stacjonarny o następujących parametrach:

|  |  |
| --- | --- |
| Procesor | AMD FX-4100 Quad Core 3.6 GHz OC |
| Płyta główna | ASRock 970 Pro 3 |
| Pamięć RAM | 12GB (2x4GB + 2x2GB Dual Channel) |
| Karta Graficzna | AMD Radeon R7 270X 2G GDDR5 |
| Zasilacz | XFX PRO450W |
| Dysk SSD | GoodRam 120GB |
| System operacyjny | Xubuntu 16.04.2 |

Zestaw ten przetestowano uprzednio pod względem wydajności zarówno dla jakości FullHD jak i UHD, stwierdzając poprawne działanie dla obu konfiguracji. Ze względów finansowych, a także trudności transportowych zrezygnowano z prób pozyskania telewizora wspierającego standard obrazu wyższy niż FullHD do testów.

Kolejnym krokiem było przygotowania stanowiska do przeprowadzenia testów. Dla testera przygotowano komfortową sofę umieszczoną centralnie naprzeciwko telewizora, umieszczonego na czarnej skrzyni. Czterdziestocalowy telewizor Samsung posiada wysokość około 50 centymetrów dlatego też umieszczono go w odległości około dwóch metrów od miejsca siedzącego testera co stanowi dystans czterech wysokości ekranu realizując założenia podane w rekomendacjach. Ponieważ pomieszczenie w którym przeprowadzano badania było dźwiękoszczelne nie należało martwić się o dochodzący z zewnątrz hałas, zadbano jednak o to aby kolejni testerzy nie przeszkadzali sobie usunięto ich z pomieszczenia w którym odbywał się test.

Sprawdzenie ilości klatek na sekundę

Weryfikacja zgodności odtwarzanych klatek z wyświetlanymi od początku pracy stanowiło problem, którego rozwiązanie wydawało się konieczne dla weryfikacji poprawności działania całego środowiska testowego. Poprawne wyświetlanie filmu jest konieczne do przeprowadzania testów. Jeżeli w trakcie odtwarzania filmu pomijane byłby losowe klatki jakość każdego z testowanych filmów w każdej sesji testu mogłaby być obiektywnie wyraźnie różna, przeprowadzanie takiego testu niemiałoby sensu ze względu na całkowicie losowe wyniki.

Rozpatrując powyższy problem postanowiono ponumerować klatki w odtwarzanej sekwencji wideo. Kolejnym problemem było umieszczenie numeracji na każdej z klatek, zrobienie tego poprzez GUI środowiska testowego okazało się bezużytecznym, ponieważ każde opóźnienie przy wczytywaniu klatek mogło spowodować desynchronizację filmu z numeracją. Zdecydowano się edytować nagranie. W tym celu posłużono wersją demonstracyjną oprogramowania firmy sony Movie Studio Platinum 13. Program udostępnia bardzo wiele opcji edycji różnego rodzaju multimediów. W tym teście kluczowa okazała się możliwość rozbicia filmu na klatki i edycji każdej z nich osobno. Na klatkach umieszczono kolejne numery. Ponumerowano około 3 sekund filmu umieszczając na nagraniu liczby od 1 do 75.



Rysunek 1 Film wczytany do Sony Movie Studio Platinum z numerowanymi klatkami



Rysunek 2 Klatka numer 6 z przerobionego filmu

Film zapisany w formacie mp4 poddano następnie dekompresji i uruchomiono przy pomocy przygotowanego oprogramowania. Obserwując odtwarzanie filmu stwierdzono wyświetlenie się wszystkich liczb co oznaczało poprawne odtwarzanie wszystkich klatek. Niestety ludzkie oko bywa zawodne, dlatego uznano, iż test należy powtórzyć nagrywając cały proces odtwarzania przy pomocy kamery pozwalającej na nagrywanie w zwolnionym tempie. Użyto w tym celu kamery smartfonu iPhone SE. Cytując za specyfikacją naukową kamera smartfonu pozwala na uruchomienie funkcji nagrywania wideo w zwolnionym tempie w jakości 1080p z częstością 120 kl./s. Nagrany film ponownie umieszczono w Movie Studio Platinum. Ponieważ oryginalny film posiadał około 25 klatek na sekundę, jego ponowne nagranie z większą częstością (120kl/s) pozwoliło na obserwację tej samej klatki kilkukrotnie. Zauważono, że każda numer został wyświetlony, dlatego też uznano, że odtwarzacz wyświetla poprawną ilość klatek.

Testy Subiektywne

Cel eksperymentu

Za cel przeprowadzanego eksperymentu przyjęto zbadanie wpływu wyboru metody przeprowadzania testu na otrzymane wyniki. Przeprowadzenie różnych testów i porównanie wyników pozwoliłoby na wyłonienie metody najbardziej efektywnej. Postanowiono zastanowić się nie tylko nad samymi wynikami, ale także nad łatwością obsługi, czasem koniecznym na przestawienie sposobu działania danego testu testerowi, opinią testerów dotyczącą poszczególnych testów oraz czasem pozyskiwania wyników z poszczególnych testów.

Wybrane filmy źródłowe

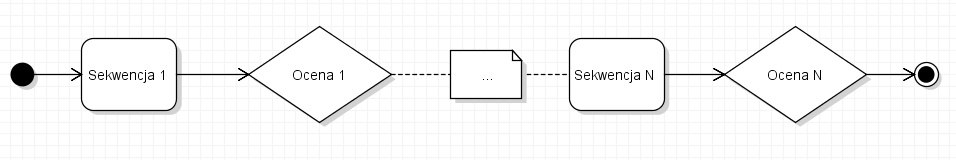
Do przeprowadzania testów subiektywnych konieczne są pliki źródłowe możliwie jak najwyższej jakości, należało pozyskać filmy udostępniane na licencji pozwalającej na ich przetwarzanie i rozpowszechnianie. Filmy pozyskano ze strony internetowej udostępniającej darmowe próbki w jakości UHD [(<http://4ksamples.com/)>]. Zdecydowano się na dwa filmy:

* PUPPIES BATH IN 4K (ULTRA HD)(Original\_H.264-AAC).mp4
* 4K-Chimei-inn-60mbps (4ksamples) .mp4

Z powyższych filmów zostały wycięte 10 sekundowe fragmenty. Pozyskane w ten sposób fragmenty nazywane odtąd odpowiednio PUPPIES i CHIMEI przetworzono przy użyciu stworzonego skryptu, opisanego w części pracy dotyczącej oprogramowania [LINK DO SKRYPTU]. Zdecydowano się utworzyć sekwencje w szerokiej rozpiętości jakości. Filmy kompresowano zmieniając przepływność rozpoczynając od 1Mb/s i kończąc na 10Mb/s przy zachowaniu stałej przepływności. Uzyskując po 10 nagrań przetworzonych dla każdego z filmów źródłowych.

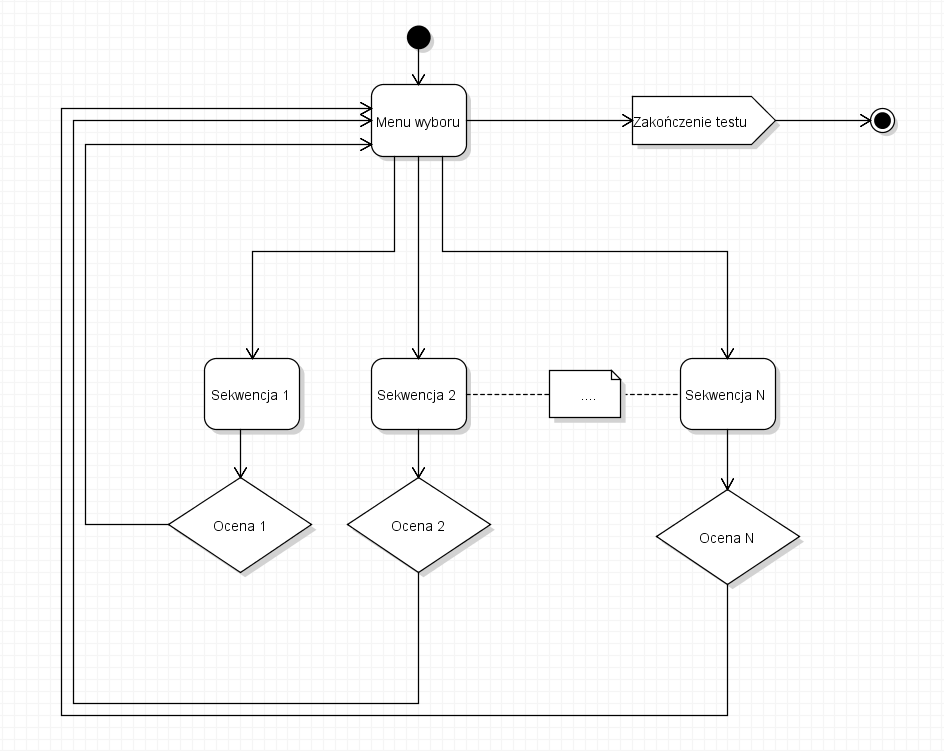
Scenariusze Testowe

Za radą promotora, a także czerpiąc wiedzę z rekomendacji ITU zdecydowano się na trzy metody przeprowadzania testów. Pierwszym z testów był przeprowadzany metodą ACR (Absolute Category Rating) polega on na wyświetlaniu filmów po kolei w losowej kolejności przy czym każdy z filmów jest oceniany natychmiast po jego odtworzeniu. Cykl przeprowadzanego testu został przedstawiony na poniższym rysunku.



Na rysunku widzimy przebieg testu od początku do końca, można zauważyć, że przebieg testu jest liniowy, tester musi obejrzeć wszystkie przygotowane sekwencje każdej z nich wystawiając ocenę, aby go zakończyć. Test ten jest najprostszym koncepcyjnie testem. Ważnym aspektem jest brak możliwości powrotu do wcześniej obejrzanego filmu, raz wystawiona ocena jest zarazem ostateczną co też może powodować, że w przypadku błędnego zaznaczenia oceny tester nie ma możliwości poprawy. Filmy odtwarzano w losowej kolejności zmieniając jakość. W teście zapytano wprost o jakość filmu. Pytanie brzmiało: „Jak oceniasz jakość obejrzanego filmu?”. Skala oceniania została oparta na rekomendacji i była skalą pięciostopniową zgodną z przykładem przytoczonym we wstępie teoretycznym [LINK]. Zdecydowano się pozostawić zarówno numery jak i oceny tekstowe (np. ocena – Bardzo dobra (5)).

Kolejną metodą przeprowadzania testu była nieopisywana w rekomendacjach metoda polegająca na udostępnieniu testerowi menu z którego ten mógł wybrać dowolny ze wszystkich dostępnych filmów obejrzeć go a następnie ocenić. Głównym założeniem tej metody była możliwość obejrzenia każdego z filmu wielokrotnie co pozwalało na zmianę oceny w wypadku pomyłki czy też uznania poprzedniej oceny z nieadekwatną po obejrzeniu filmu w innej jakości. Zarys przebiegu testu przedstawiono na poniższym rysunku.

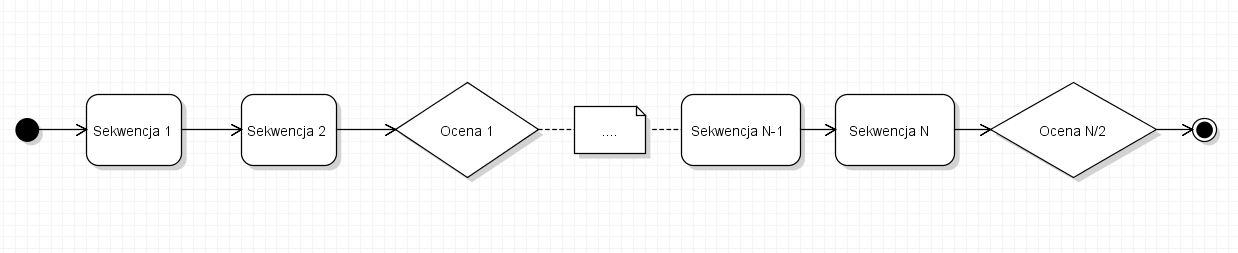


Zgodnie ze schematem przebieg testu nie ma charakteru liniowego użytkownik sam wybiera film z listy dostępnych po wystawieniu oceny trafiając ponownie do tego samego menu. Interfejs użytkownika został zaprojektowany w taki sposób, aby użytkownik każdorazowo mógł zobaczyć własne oceny.

[screen rozwiniętej listy z wystawionymi ocenami]

Jak widać na powyższym zrzucie ekranu wykonanym podczas przeprowadzania testu filmy, które zostały już ocenione posiadają oceny inne niż zero. W metodzie zostały użyte dokładnie te same nagrania których używano w teście ACR, spodziewano się więc uzyskania podobnych wyników różniących się tylko w przypadku pojedynczych filmów. Zadano to samo pytanie, pozostawiając testerowi taką samą skale oceniania.

Trzeci scenariusz został oparty o standardową metodę porównania parami (PC - Pair Comparision) polegającą na zestawianiu kolejnych filmów parami pytając o jakość drugiego z filmów względem pierwszego. W tym teście zadano pytanie: „Jak zmieniła się jakość oglądanego filmu?”. Korzystając ze skali ocen z zalecanej dla testów CCR/PC z rekomendacji, która została pokazana we wstępie teoretycznym [link] uzyskano dane do analizy opisanej w dalszej części filmu. Przebieg testu został przedstawiony na poniższym rysunku.



Testerzy

Testerów pozyskiwano wśród znajomych i rodziny. Ze względu na konieczność dojazdu do miejsca przeprowadzania testów, a także ich długość (testerów proszono o zarezerwowanie sobie około godziny) okazało się to wyjątkowo trudne. Udało się uzyskać pełne wyniki od trzynastu osób. Starano się uzyskać jak najbardziej zróżnicowaną grupę testerów. Testerzy pochodzili z różnych grup wiekowych jednak większość z nich to osoby poniżej dwudziestego piątego roku życia, część testerów była osobami posiadającymi różne wady wzorku, zachowano równy podział według płci (7 mężczyzn i 6 kobiet). Ze względu na niewielką ilość testerów nie dzielono ich na poszczególne grupy (np. według wieku, czy płci) tylko traktowano jako jedna populacja. Głównym celem badań było porównanie metod przeprowadzania testów, dlatego też nie było konieczności grupowania testerów.

Przebieg testów i obserwacje

Testy odbywały się w przygotowanym wcześniej pomieszczeniu, a przed rozpoczęciem testu zapewniono maksymalną wygodę każdemu z testerów dostosowując w miarę możliwości ustawienia myszy i klawiatury komputera, ponieważ oceny były wprowadzane korzystając z interfejsu ekranowego konieczne było zastosowanie urządzeń wejściowych. Każdy z trzech scenariuszy był uruchamiany jako osobny test, aby wyraźnie zasygnalizować zmianę metody, a także pozwolić na zadanie pytań i udzielić ewentualnych odpowiedzi co do strony użytkowej działania interfejsu odpowiedzialnego za daną metodę. Idealnym byłoby uruchamianie testów z możliwie jak największym odstępem pomiędzy kolejnymi scenariuszami jednakże ze względu na ograniczenie wynikające z czasu i cierpliwości testerów konieczne było przeprowadzenie wszystkich testów jednocześnie. Niestety mogło to mieć bardzo duży wpływ na wyniki. Jeżeli testerowi pozwolono by zapomnieć wykonywany test i obejrzane filmy jego ocena nie byłaby obciążona poszukiwaniem konkretnego filmu, który zapadł mu w pamięci jako ten z wyraźnie lepszą, bądź wyraźnie gorszą jakością. Poniższa tabela przedstawia szacowane długości poszczególnych testów.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Numer testu | Ilość filmów | Łączny czas filmów | Czas na ocenę jednego filmu | Łączny czas testu | Ilość wyników |
| 1 | 23 | 3 minuty i 50 sekund | ~10 sekund | ~8 minut | 23 |
| 2 | 23 | 3 minuty i 50 sekund | ~10 sekund + czas wyboru kolejnego filmu ~10 sekund | ~12 minut | 23 |
| 3 | 30 | 5 minut | ~15 sekund | ~9 minut | 15 |

W ostatnim z wykonywanych testów celowo zamieszczono więcej sekwencji aby uzyskać podobny czas testu, a także aby uzyskać więcej wyników uzyskując możliwość lepszego porównania poszczególnych testów. Czas na ocenę został oszacowany z obserwacji oraz pomiarów łącznego czasu trwania testów dla kilku testerów. Zauważono duże różnicę pomiędzy czasem oceniania początkowych filmów, a czasem oceniania filmów z końca danego testu. Było to spowodowane przyzwyczajeniem się użytkownika do korzystania z interfejsu w miarę wykonywania testu, a także coraz większym znudzeniem co wpływało nie tylko na chęć jak najszybszego zakończenia testu, ale prawdopodobnie także na nonszalancję w sposobie oceniania. Ze względu na wygodę testerów nie chciano zmuszać ich do czekania po wystawieniu oceny każdego filmu dlatego nie wprowadzano minimalnego czasu trwania oceny. Czas spędzony nad oceną został wydłużony w przypadku scenariusza drugiego ze względu na konieczność wybrania filmu z listy, wzięto także poprawkę na ewentualne powtórne oglądanie tego samego filmu celem zmiany oceny. Ze względu na brak kontroli nad testerem przez cały przeprowadzany test nie pozwolono na zmianę wystawionej oceny bez uprzedniego obejrzenia filmów. W teście numer 3 wydłużono szacunkowy czas oceny ze względu na konieczność porównania obu obejrzanych filmów.

Całość badania trwała w przypadku każdego z testerów około 35 minut (wraz z przygotowaniem stanowiska i krótkim instruktażem) co w przypadku kilkuosobowej grupy testerów powodowało wydłużenie czasu oczekiwania, co mogło mieć negatywny wpływ na nastawienie niektórych testerów do samego testu. Już po wykonaniu dwóch trzecich badania zauważano wyraźnie zniechęcenie każdej z osób do kontynuacji testów, co sugerowałoby konieczność zastosowania większej ilości plików źródłowych zwłaszcza dla dłuższych testów.

Testy 1 i 3 miały podobny czas trwania, zgodnie z oczekiwaniami test pozwalający na wybór filmu okazał się wyraźnie dłuższy. Mimo to test trzeci zgodnie z rekomendacją dostarcza mniejszej ilości wyników. Powodem jest oczywiście wystawianie ocen dla dwóch filmów jednocześnie. Jednakże zgodnie z rekomendacją wyników powinno być mniej o nieco mniej niż połowę w tym samym czasie testu. W zbliżonym czasie trwania uzyskano tylko około 30% mniej wyników. Należy jednak zwrócić uwagę na to iż test był dosyć krótki, a czas konieczny na wystawienie oceny coraz mniejszy w miarę upływu czasu, dlatego też można wnioskować, że dla dłuższego testu strata wyników byłaby większa.

Opinie testerów

Po przeprowadzeniu testów zadbano o zebranie opinii testerów na temat całości badań. Zauważono duże rozbieżności w opiniach. Należy zwrócić uwagę na fakt, iż większość testerów to nie osoby bezpośrednio zainteresowane tematyką QoE, a są to zwyczajni konsumenci. Testerzy zwracali uwagę na:

* Dużą rozbieżność w „trudności” oceny jakości. Niektóre z filmów były wyraźnie gorszej jakości co dawało podstawy do bardzo niskich ocen. Z kolei niektóre zdawały się niczym nie różnić
* W przypadku testu z menu wyboru, pierwszy kontakt z listą sekwencji wydawał się przytłaczający
* Niektórzy testerzy uznali, że możliwość powrotu nie ma sensu z kolei inni chętnie z niej korzystali, zwłaszcza w początkowej fazie testu
* Film „Chimei” w początkowej części miał zdecydowanie gorszą jakość niż w końcowej
* Duża powtarzalność i monotonia badań

Ciekawym są również różne wskazania testerów w kwestii najlepszego ich zdaniem testów. Niektórzy uznali pierwszy test za najlepszy wskazując na jego prostotę i szybkość. Jeden z testerów zajmujący się zawodowo prowadzeniem testów automatycznych wskazał także dużą ilość danych dostarczanych przez pierwszy ten w stosunku do jego czasu trwania. Krytykowano jednak brak możliwości powrotu i zbyt wąską skale ocenienia. Część testerów za najlepszy uznała test trzeci, jako najciekawszy ze względu na brak konieczności myślenia o obejrzanych wszystkich dotychczas filmach, lecz możliwości skupienia się tylko na dwóch.

Ogólne wnioski

Główną wadą przeprowadzonych badań była ich powtarzalność, różnice miedzy filmami kompresowanych z użyciem wyższej przepływności były dla większości testerów niezauważalne, dlatego też skarżyli się oni na to że oceniane sekwencje są takie same przez co ocenianie jest trudne, a sam test jest nudny. Zwrócono jednak uwagę na fakt, iż w warunkach komercyjnych przeprowadzanie testów subiektywnych dla filmów o niskiej jakości nie ma sensu, ponieważ każdemu producentowi sprzętu czy usługodawcy zależy na dostarczeniu jak najwyższej jakości, dlatego też czułość metody testowej jest bardzo ważna.

Testy subiektywne poza danymi dotyczącymi subiektywnej oceny (subiektywnego współczynnika jakości – MOS) dostarczają bardzo wielu danych odnośnie psychologii i teorii podejmowania decyzji. Zauważono, że testerzy w miarę przebiegu testu starali się myśleć o wszystkich filmach jednocześnie starając się odnosić oceny poszczególnych filmów do wcześniej obejrzanych, mimo poinstruowania ich, aby każdą z sekwencji traktować osobno. Następowała więc silna relatywizacja opinii o każdym kolejnym z filmów. Ciekawa jest także rozbieżność w ocenach sekwencji w tej samej jakości w różny sposób w zależności od poprzedzających filmów.