DOI: 10.51480/compress.2021.4-1.248

Dominika Saad
SWPS UNIWERSYTET HUMANISTOSPOŁECZNY
0000-0003-4170-7013
saad.dominika@gmail.com

# Nowe narzędzia i techniki zwiększające trafność badań internetowych

## Increasing Validity of Online Research by Implementing New Tools and Techniques

**ABSTRAKT**

Celem artykułu jest przedstawienie nowych technik poprawiających jakość danych uzyskiwanych w badaniach przeprowadzanych online na przykładzie panelu Amazon MTurk. Poprzedzona kwerendą, krytyczna analiza literatury przedmiotu identyfikuje główne źródła zniekształcenia wyników, którymi są: bezrefleksyjność, działalność botów wypełniających ankiety oraz zachowania respondentów, klasyfikowane jako nadużycia lokalizacyjne interfejsu sieciowego IP. Wykorzystane podczas badania narzędzia oraz techniki wskazują na praktyczne sposoby zwiększania trafności uzyskiwanych danych poprzez rozpoznanie wyników zakłócających na poziomie 60%. Analiza otrzymanych danych potwierdza: 1) użyteczność zastosowanych technik kontroli oraz 2) stosowalność zaproponowanego w tekście modelu operacjonalizacji czasu jako istotnego wskaźnika zachowań bezrefleksyjnych.

**SŁOWA KLUCZOWE: BEZREFLEKSYJNOŚĆ, NADUŻYCIA IP, AMAZON MTURK, DANE NISKIEJ UŻYTECZNOŚCI, ZWIĘKSZENIE TRAFNOŚCI**

**ABSTRACT**

The aim of this paper is to present new techniques that increase the quality of data obtained through online research, basing on the example of Amazon MTurk. Using a critical analysis of the literature, preceded by a query, we identified the main sources of low-quality data such as careless responses, form-fill bots activity, and fraudulent behavior of manipulating IP addresses. The techniques implemented in the study offer several practical implications in terms of validity improvement. The results proved that 60% of observations were recognized as poor quality data. This confirms the applied control techniques to be useful and the proposed theoretical model of time operationalization to be an important indicator of careless responses.

**KEYWORDS: CARELESS RESPONSES, FRAUDULENT IP MANIPULATION, AMAZON MTURK, POOR QUALITY DATA, VALIDITY IMPROVEMENT**

Saad, D. (2021), Nowe narzędzia i techniki zwiększające trafność badań internetowych, *com.press*, 4(1), s. 106–121. DOI: 10.51480/compress.2021.4-1.248 www.compress.edu.pl

## WSTĘP

W związku z wprowadzeniem izolacji społecznej jako jednego ze środków zapobiegania rozprzestrzenianiu się pandemii COVID-19 badacze postawieni zostali wobec konieczności reorganizacji wielu podejmowanych przez siebie przedsięwzięć i dostosowania się do nowych, wciąż pojawiających się ograniczeń. Dystans społeczny wymusił porzucenie, przełożenie bądź modyfikowanie projektów badawczych wiążących się z bezpośrednimi interakcjami, podróżami czy nawet opuszczeniem miejsca zamieszkania. W rezultacie uczeni wielu dziedzin zmuszeni zostali do skorzystania z ograniczonej palety narzędzi badawczych pozwalających na zdalne przeprowadzanie badań. Trafność otrzymanych wyników może być zniekształcona ze względu na sposób, w jaki respondenci reagują na przeprowadzane manipulacje eksperymentalne online. Celem pracy jest podsumowanie wiedzy dotyczącej potencjalnych źródeł zagrożeń trafności otrzymywanych wyników, omówienie nowych technik przeprowadzania eksperymentów sondażowych online oraz przedstawienie praktycznych sposobów rozpoznawania i wykluczania danych zakłócających.

Zastosowane w tekście metody badawcze to – poprzedzona kwerendą – krytyczna analiza literatury przedmiotu, publikacji naukowych dostępnych za pośrednictwem zbiorów bazy Google Scholar oraz empiryczne badania przeprowadzone online w terminie październik–grudzień 2020 r. Badanie zrealizowane zostało za pomocą panelu Amazon Mechanical Turk. Do wypełniania kwestionariusza przystąpiło 342 respondentów. W instrukcji podano informacje o wymaganiu wiekowym i wystąpieniu kontroli uwagi. Zawarta została również informacja, że warunkiem otrzymania wynagrodzenia jest udzielenie co najmniej 66% poprawnych odpowiedzi na pytania kontrolujące uwagę.

Kwestionariusz w angielskiej wersji językowej, składający się ze 111 pytań i metryczki, wprowadzony został za pomocą narzędzia Qualtrics. Proces ten pozwolił na zwiększenie kontroli poprzez automatyczne wstawianie reCaptcha oraz wygenerowanie unikalnego, losowego numeru ułatwiającego system rozliczeń wynagrodzeń dla respondentów platformy ATM. Qualtrics umożliwia stworzenie logiki przebiegu schematu badania. Pozwala to na weryfikację warunków w trakcie wypełniania kwestionariusza. W badaniu zastosowano dodatkowy test – równanie antybotowe samodzielnego autorstwa[1] oraz weryfikację deklarowanego wieku respondentów. Za pomocą wprowadzonej

1 Przykład równania antybotowego wraz z instrukcją znajduje się w kolejnej części tekstu.

logiki odpowiedzi respondentów były zapisywane, wysyłane na serwer i odsyłane, generując odpowiednią reakcję, czyli przejście do dalszej części badania bądź jego zakończenie. Warunkiem dostępu do kolejnej części badania było przejście co najmniej jednej z dwóch kontroli antybotowych oraz jednoczesne wpisanie w rubrykę wieku liczby równej lub większej od 55. Niespełnienie powyższych założeń odsyłało badanego do ekranu kończącego badanie. Jakość zebranych danych została zweryfikowana przy uwzględnieniu głównego źródła zniekształceń opisywanego w literaturze przedmiotu. Kontrolowano uwagę, aktywność botów, ryzyko związane z nadużyciem fałszowania lokalizacji, czas wypełniania kwestionariusza oraz wielokrotne wypełnienie kwestionariusza przez respondentów.

## ŹRÓDŁA ORAZ TECHNIKI IDENTYFIKUJĄCE DANE NISKIEJ UŻYTECZNOŚCI

Jak pokazują statystyki Google Scholar, dane pozyskane za pomocą badań przeprowadzonych online są powszechnie wykorzystywane w procesach naukowych. W 2020 roku Google Scholar w swoich zasobach cytuje ponad 8 tysięcy publikacji z użyciem frazy „Amazon Mechanical Turk" (AMT)[2], najpopularniejszej strony crowdsourcingowej założonej w 2005 r. Platforma AMT używana jest między innymi przez badaczy do zlecania Human Intelligence Task (HIT) w postaci zadania, np. wypełnienia ankiety w zamian za wynagrodzenie. Pomiędzy rokiem 2015 a 2020 AMT był cytowany 24 tysięcy razy przez Google Scholar. Jak pokazują badania (Becker i in., 2013) pomimo selekcji niektórych grup respondentów (im starsi użytkownicy, tym rzadziej korzystają z urządzeń mobilnych jako preferowanego narzędzia do wypełnienia ankiety) kwestionariusze wyświetlane na urządzeniu mobilnym (smartfon, tablet) były stosowane już ponad dekadę temu, a ich użyteczność z punktu widzenia badawczego wielokrotnie sprawdzono empirycznie (Tourangeau i in., 2017). Kees i inni (2017) wskazują wyższą jakość wyników otrzymywanych podczas badań przeprowadzonych na AMT nad panelami internetowymi oraz porównywalnymi wynikami pozyskanymi z badania populacji studenckich. Hargittai (2020) przedstawia odmienne wnioski, wskazując istotne różnice w kompetencjach i cechach socjodemograficznych

2 W Polsce istnieją podobne panele, np. Panel Ariadna, Imas, Survey Compare, panel GFK Polonia, poznaj to, gdzie za wypełnianie ankiet otrzymać można gotówkę lub vouchery.

respondentów ATM, które – podobnie jak w przypadku grup studenckich – mogą wpływać na tzw. błąd próby.

Badania prowadzone za pośrednictwem internetu mają niewątpliwie wiele zalet. Są one relatywnie tanie, dostępne, umożliwiają interaktywność poprzez łatwe wykorzystanie multimediów, pozwalają na dotarcie do dużej liczby specyficznych respondentów w krótkim czasie, ułatwiają zbieranie i zapisywanie danych oraz charakteryzują się dużą elastycznością. Pomimo wielu wymienionych wyżej zalet prowadzenie badań w przestrzeni internetowej ma również istotne wady. Fizyczna odległość badacza od respondentów sprzyja brakowi kontroli i utrudnia weryfikację czynników mogących mieć wpływ na otrzymane wyniki. W drugiej połowie 2018 r. w mediach społecznościowych pojawiły się doniesienia o niereplikowalnych wynikach badań psychologicznych[3], danych niskiej jakości i użyteczności, potencjalnych botach oraz algorytmach wypełniających kwestionariusze online. Zapoczątkowało to falę dyskusji o użyteczności AMT i idące za nią liczne badania mające na celu zdiagnozowanie przyczyn leżących u źródła zakłóceń otrzymywanych wyników. W badaniu (Moss, 2018) testowano hipotezy dotyczące zniekształceń dokonywanych przez potencjalne boty wypełniające, „bezrefleksyjne", czyli nieuważne lub losowe wypełnianie ankiet przez respondentów. Moss dokonał również próby diagnozy pochodzenia użytkowników AMT dostarczających dane niskiej jakości. Zgodnie z wstępnymi założeniami wykazano, że boty stanowią relatywnie małe i dość proste do wyeliminowania zagrożenie dla eksperymentów prowadzonych online (o czym w dalszej części tekstu). Większość narzędzi lub platform badawczych umożliwia wstawianie mechanizmów weryfikacji antybotowych, takich jak test reCaptcha, bądź wstawia je automatycznie, wychwytując boty już przed rozpoczęciem wypełniania ankiety. ReCaptcha to prosty test, który ze względu na specyficzną konstrukcję jest intuicyjny i prosty do rozwiązania dla człowieka, lecz relatywnie trudny dla bota. Zadanie antybotowe polega zazwyczaj na nazwaniu lub wyodrębnieniu obrazów przedstawiających ten sam obiekt na kilka różnych sposobów. Po zastosowaniu mechanizmów blokujących boty wykazano, iż głównym źródłem zniekształcenia są niskiej

3 Kwestia niereplikowalnych badań psychologicznych w odniesieniu do klasycznie przeprowadzonych eksperymentów i jej ważność podniesiona została już wcześniej przez Everetta Trafimowa (2015). Informacje o publikacji znajdują się w bibliografii.

jakości dane wynikające z czynnika ludzkiego (Moss, 2018)[4]. Zniekształcenia i nieefektywność zastosowanych eksperymentalnie manipulacji wynikała głównie z bezrefleksyjności odpowiedzi oraz niskiego poziomu znajomości języka angielskiego, uniemożliwiającej zrozumienie instrukcji zamieszczonych w badaniu. Analiza pochodzenia danych wykazała, że większość niskiej jakości wyników pozyskana została przez respondentów pochodzących głównie z Wenezueli i Indii, mimo iż warunkiem udziału w badaniu było połączenie się adresem IP rodzimym, zlokalizowanym na terenie Stanów Zjednoczonych[5]. Istnieją jednak relatywnie proste sposoby zablokowania widoczności IP lub połączenia się przy użyciu farmy serwerowej tzw. VPN, która pozwala na maskowanie lokalizacji lub fałszuje ją. Umożliwia to osobom spoza preferowanej próby dostęp do badania oraz otrzymanie rekompensaty za udział. To zdaje się jest kolejny powód otrzymywania zniekształconych wyników w wielu badaniach przeprowadzanych za pośrednictwem platformy AMT. Respondenci łączący się poprzez farmy serwerowe dostarczają danych o niskiej użyteczności, częściej odpowiadają bezrefleksyjnie i prawdopodobnie z powodu niskiego poziomu zrozumienia, zaangażowania i motywacji reagują znacznie słabiej na warunki manipulacji eksperymentalnej nawet dobrze znanych i replikowanych eksperymentów, jak dylemat wagonika i efekt zakotwiczenia (Moss, 2018)[6]. W badaniu przeprowadzonym w 2020 r. przez zespół K. Peytona sprawdzano trafność zewnętrzną trzynastu badań w trzydziestu trzech replikacjach, próbując ustalić, czy pandemia wpływa na zachowania respondentów online. Otrzymane wyniki nie wykazały różnic w zakresie widocznych zmian badanych zmiennych, pokazały natomiast osłabienie efektu, czyli zmniejszenie trafności wewnętrznej. Badacze wskazują, że wpływ na to zjawisko ma mała uważność. Wnioski z badań zdają się spójne w odniesieniu do bezrefleksyjności oraz niezweryfikowanego pochodzenia respondentów generujących dane o niskiej jakości jako główne źródła zakłóceń badań przeprowadzanych online. Dla kontrolowania trafności wskazane jest zatem uwzględnianie wyżej wymienionych czynników podczas procesu badawczego.

4 W badaniu zastosowano formularz, który w przypadku bota, czyli zautomatyzowanego skryptu, skutkował zaznaczeniem wszystkich pól odpowiedzi. Boty i algorytmy wypełniające programowane są tak, aby automatycznie wypełniać treści. Tak proste mechanizmy nie posiadają możliwości rozumienia poleceń, np. „Wypełnij jedno pole z dwóch". Na tej podstawie relatywnie łatwo rozpoznać bota.

5 IP to liczba nadawana interfejsowi (łączu) sieciowemu służąca identyfikacji w obrębie sieci lokalnej lub poza nią.

6 W badaniu wykazano istotnie słabszą siłę efektu.

Bezrefleksyjność, zwana również „przeklikiwaniem" ankiet, to powszechnie opisywany problem prowadzący do otrzymywania niereprezentatywnych wyników. Tourangeau (2000) opisuje cztery etapy składające się ze specyficznych procesów poznawczych, które występują podczas odpowiadania na pytanie ankietowe. Etapy te przebiegają kolejno i są to: zrozumienie (obejmuje wymóg procesu zrozumienia i odpowiedzi na pytanie zgodnie z zamieszczoną instrukcją), wyszukiwanie (dotyczy procesu poszukiwania odpowiednich informacji dotyczących pytania), osąd (proces decyzji uwzględniający wyszukane informacje), odpowiedź (przyporządkowanie decyzji do wybranej odpowiedzi). Do prawidłowego przetworzenia informacji i przebiegu wymienionych procesów niezbędny jest czas oraz pewien poziom zaangażowania uwagi respondenta. Na ilość wysiłku, który podejmie badany, wpływ ma jego motywacja. Przyczyną bezrefleksyjnego odpowiadania, przypadkowego zaznaczania lub niespójnych odpowiedzi może być zatem niewystarczający poziom motywacji badanych. Jest ona niezbędna, aby poprawnie zinterpretować zawartość pozycji, zastosować się do instrukcji ankiety i udzielić adekwatnej odpowiedzi (Huang i in., 2012). Zasadne zdaje się odniesienie do kontekstu adresowanego problemu. Uczestnicy badań paneli oferujących wynagrodzenie mogą być jednocześnie wysoce zmotywowani i bezrefleksyjni. Ich motywacją może być jak najszybsze – zamiast np. skrupulatne bądź uważne – wypełnienie kwestionariusza.

Jak pokazuje przegląd badań empirycznych, poziom zarejestrowanych bezrefleksyjnych odpowiedzi wynosi nawet 78% (Mancosu i in., 2019)[7] w różnych tematykach i kontekstach badawczych, stanowiąc poważne zagrożenie dla kryterium ważności miar. Wyniki badań wskazują konieczność stosowania skutecznych oraz dobrze dopasowanych technik kontroli uwagi respondentów. W literaturze przedmiotu pytania monitorujące bezrefleksyjność respondentów nazywane są często *attention check*, czyli kontrolą uwagi (KU). W eksperymencie badającym własną percepcję jakości wypełniania formularzy online przez respondentów z platformy AMT (Lovet i in., 2018) aż 70% badanych deklarowało, że wypełnia ankiety na „bardzo wysokim poziomie", a pozostałe 30% oceniło jakość swojej pracy na „wysoki poziom". Co interesujące, ani jedna osoba nie wartościowała swojej pracy „poniżej przeciętnej". Ponadto 55% respondentów tego samego panelu deklarowała, że „bardzo często" aktywnie szuka pytań kontrolujących uwagę w ankietach,

7 W eksperymencie badacze sprawdzali uważność respondentów, podając im niespójne informacje, a następnie prosili o zaznaczenie konkretnych odpowiedzi, manipulując obciążeniem poznawczym (np. tekstem zawierającym trudne wyrazy). Bezrefleksyjność mierzono za pomocą liczby poprawnych odpowiedzi.

37% „raczej często", a jedynie 8% zdaje się nie mieć wiedzy o występowaniu KU w kwestionariuszach. Wyraźnie widać zatem, że wzrost popularności badań internetowych zmienił zdecydowanie świadomość i zachowania respondentów. Badani postrzegają więc własny udział jako charakteryzujący się wysokim poziomem użyteczności, a zarazem – pomimo wyraźnego poszukiwania kontroli uwagi – systematycznie i często odpowiadają bezrefleksyjnie.

W uprzednio cytowanych badaniach uznaje się, że bezrefleksyjność jest mierzalna i może być skutecznie kontrolowana poprzez wstawianie nieprawdziwych lub nieprawdopodobnych stwierdzeń do kwestionariusza, mierzenie częstości zaznaczania odpowiedzi pułapek przez respondentów i wykluczanie zniekształcających danych w analizie *post hoc*[8]. W badaniu (Meade i in., 2012) przeprowadzonym na grupie studentów psychologii zastosowano stwierdzenia wplecione w różnego rodzaju kwestionariusze, takie jak: „wszyscy moi znajomi są kosmitami", „co dwa tygodnie płacą mi skrzaty" czy „sypiam mniej niż godzinę dziennie". Uzyskane wyniki wskazały, iż 10–12% respondentów wykazała się bezrefleksyjnością, zaznaczając „zdecydowanie się zgadzam" na pytania pułapki badaczy. Istnieją również inne formy stosowania technik sprawdzających poziom bezrefleksyjności. Jedną z nich jest wstawianie dodatkowego pytania KU z wyraźną instrukcją w jego bezpośredniej treści o pozostawieniu go bez odpowiedzi (np. „Uważam, że jestem osobą otwartą na innych. Proszę, pozostaw to pytanie bez odpowiedzi/nie zaznaczaj żadnej odpowiedzi w tym pytaniu") lub pytanie z wyraźnie wskazaną prośbą konkretnej reakcji (np. „Uważam, że jestem osobą otwartą na innych. Proszę, zaznacz w tym pytaniu odpowiedź: zdecydowanie się nie zgadzam"). Pytania z instrukcją o pozostawieniu pozycji bez jakiejkolwiek reakcji wydają się skuteczniejsze, ponieważ przy konstrukcji formularza z opcją wymuszenia odpowiedzi w każdej pozycji, oprócz pytania kontroli uwagi KU w tej formie, zaznaczenie przez respondenta jakiejkolwiek odpowiedzi świadczy o bezrefleksyjnej odpowiedzi i może być podstawą do wykluczenia jego wyników z analizy. W pytaniu KU z instrukcją zaznaczającą wskazanie konkretnej odpowiedzi istnieje szansa losowego przejścia kontroli. Udane zabiegi mające na celu wykluczenie bezrefleksyjnych odpowiedzi są ważnym etapem zwiększającym trafność otrzymanych wyników, a stosowanie ich nie wykazało uszczerbku dla trafności pomiarów (Kung, 2018). Stosując metody kontroli typu KU, warto również poinformować o fakcie sprawdzania

8 Istnieją również inne stosowane sposoby pomiaru uważności, polegające na analizie treści pytań otwartych, mierzeniu czasu wykonania kwestionariusza lub częstości zaznaczania tej samej odpowiedzi z rzędu. Więcej informacji dostępnych jest w załączonym spisie bibliograficznym.

uwagi respondentów w trakcie badania już w instrukcji, co może zwiększyć uwagę badanych. Pytania typu KU rekomenduje się stosować z umiarem, wykorzystując randomizację. Badanie Mead i in., (2012) wskazuje, że pytanie kontrolujące uwagę przynosi optymalne rezultaty w proporcji 1–2 na 50–100 pytań kwestionariuszowych.

Oprócz omówionych wyżej technik kontroli bezrefleksyjności należy zwrócić uwagę także na czas wypełnienia kwestionariusza przez badanego. Bardzo krótki czas ukończenia może również świadczyć o bezrefleksyjnych zachowaniach. Pewnego rodzaju trudność stanowi oszacowanie adekwatnej ilości czasu potrzebnego do wypełnienia ankiety. Wynika to między innymi z różnic indywidualnych w prędkości czytania. Można jednak przyjąć pewne kryteria operacjonalizacji czasu. Bell (2001, za: De Leeuw, 1965; Fry, 1963) proponuje podział na osoby czytające szybko, umiarkowanie i wolno, to jest odpowiednio 350, 250 i 150 słów na minutę. Poprzez zliczenie wszystkich wyrazów kwestionariusza i podzielenie ich przez liczbę słów dla czytających w różnym tempie grup otrzymamy trzy wyniki stanowiące punkt odniesienia. Czas przeczytania ankiety nie jest jednak równoznaczny z czasem potrzebnym na jej wypełnienie. Należy uwzględnić również „zapas" niezbędny na zaistnienie wymienionych wcześniej procesów zrozumienia, wyszukania, osądu oraz zaznaczenia odpowiedzi (Tourangeu, 2000). Zasadne jest zatem logicznie przyjąć, iż czas wypełnienia ankiety jest dłuższy niż czas potrzebny na jej przeczytanie[9].

Wyniki przytoczonych badań wskazują sposoby na wywołanie istotnie skutecznego wpływu poprzez stosowanie odpowiednich technik i narzędzi umożliwiających modelowanie bezrefleksyjnych zachowań respondentów. Relatywnie proste do wykonania zabiegi pozwalają również na zwiększenie szansy na szybkie „wyłowienie" na wczesnym etapie analizy nadużycia respondentów zniekształcających otrzymane wyniki.

9 Założenie to zostało również przyjęte i zastosowane podczas analizowania danych zebranych w badaniu własnym. W dalszej części przedstawiono dokładny sposób operacjonalizacji zmiennej czasu oraz oszacowanie, jakie wyniki mogą świadczyć o bezrefleksyjności.

Dominika Saad
SWPS Uniwersytet Humanistospołeczny

# ANALIZA TECHNIK ZASTOSOWANYCH W BADANIU WŁASNYM PRZY *N* POCZĄTKOWYM NA POZIOMIE 342 OBSERWACJI

## AUTOMATYCZNA KONTROLA ANTYBOTOWA RECAPTCHA

reCaptcha zastosowanego narzędzia Qualtrics zidentyfikowała i wykluczyła $n = 17$ potencjalnych botów, wyniki te nie zostały uwzględnione w dalszej analizie.

## KONTROLA ANTYBOTOWA – RÓWNANIE

Instrukcja wyświetlona respondentom przed pytaniami zawierała prośbę badaczy o rozwiązanie prostego działania matematycznego zapisanego słownie i wpisanie wyniku również słowami przy użyciu wyłącznie wielkich liter[10]. Poprawna odpowiedź na przynajmniej jedno zadanie umożliwiała przejście do kolejnej części kwestionariusza. W przypadku poprawnej odpowiedzi na pierwsze pytanie uczestnik był odsyłany do części weryfikującej wiek. W razie niepowodzenia respondentowi wyświetlano drugie pytanie – równanie. Nieprawidłowe odpowiedzi na obydwa pytania zdyskwalifikowały 7,31% ($n = 25$) wyników. Pytania skonstruowane z myślą o kontrolowaniu botów dostarczyły dodatkowych, niespodziewanych danych. Wiele błędnych odpowiedzi sugerowało znajomość języka angielskiego na bardzo niskim poziomie. Niezrozumienie zdaje się być przyczyną błędnych odpowiedzi. Pojawiały się odpowiedzi świadczące o tym, że respondenci nie rozumieli podstawowych znaków działań. Drugie pytanie (ile to jest cztery podzielić na dwa?) generowało często odpowiedzi: sześć lub osiem. Sugeruje to obecność obcojęzycznych respondentów o niskich kompetencjach w zakresie języka angielskiego. Błędne rozumienie tego typu prostego działania rodzi wątpliwości co do zdolności do wypełnienia kwestionariusza na poziomie umożliwiającym zrozumienie i prawdziwe ustosunkowanie się do jego instrukcji i pytań.

10 Przykład równania antybotowego: „Proszę, rozwiąż poniższe działanie. Wpisz swoją odpowiedź słownie, używając wyłącznie wielkich liter. Ile to jest trzy plus siedem? Ile to jest cztery podzielić na dwa?”.

### WERYFIKACJA WIEKU RESPONDENTÓW

Celem badawczym była również weryfikacja możliwości dotarcia do specyficznej grupy najstarszych respondentów. Korzystanie z ATM umożliwia otwarcie panelu dla grupy respondentów 55+. Weryfikacja polegała na wpisaniu w rubrykę z wiekiem liczby równej lub większej niż 55. Niespełnienie powyższego założenia odsyłało badanego do ekranu kończącego badanie. Odnotowano liczne próby nadużycia w postaci wpisywania przez tego samego respondenta różnych parametrów wieku z innego IP. Pewien rezolutny uczestnik zainkasował kilkakrotną wypłatę środków za udział w badaniu, obchodząc logikę warunku i wpisując swój wiek w formie liter (dwadzieścia sześć), a nie cyfr – co weryfikował algorytm. Niespełnione kryterium wieku wykluczyło 10,2% ($n$ = 35) obserwacji. Dostęp młodszych uczestników do badania pokazuje również niedoskonałość platformy ATM i możliwe częste nadużycia respondentów w tym zakresie.

### WYŁAMANIE SIĘ

W trakcie wypełniania kwestionariusza 7,02% ($n$ = 24) uczestników zrezygnowało z dalszego wypełniania. Obserwacje zostały odrzucone.

### KONTROLA BEZREFLEKSYJNOŚCI

W badaniu zastosowano trzy kontrole uwagi typu pytania KU, wplecione w pytania kwestionariusza i pojawiające się w losowej kolejności. Dwie pozycje KU zawierały w treści polecenie zaznaczenia konkretnej odpowiedzi skrajnej na skali. Trzecia kontrola uwagi prosiła respondentów o pozostawienie pytania bez odpowiedzi. Było to jedyne pytanie w badaniu, które można było pozostawić bez zaznaczenia. Wszystkie pozostałe pozycje zawierały warunek wymuszający odpowiedź na każde pytanie w celu ukończenia kwestionariusza i otrzymania kompensaty[11]. Zaznaczenie jakiejkolwiek pozycji było zatem jednoznaczne z bezrefleksyjnością. 69,4% respondentów ($n$ = 179) prawidłowo zareagowało na wszystkie pytania, 10,1% ($n$ = 26) poprawnie

11 W nawiązaniu do etycznej kwestii badań należy wskazać konieczność poinformowania respondentów przed przystąpieniem do badania, iż kompensata będzie wypłacana jedynie w przypadku podania odpowiedzi na wszystkie pytania. Aspekt ten jest istotny ze względu na kwestię udzielania przez uczestnika świadomej zgody na wspomniany warunek, przez który rozumieć należy fakt rozpoczęcia udziału w badaniu. Warto również zaznaczyć, że w sytuacji, w której odpowiedź nie jest dla respondenta poznawczo dostępna, istnieje prawdopodobieństwo, że udzieli odpowiedzi losowej bądź zaprzestanie udziału w badaniu. W sytuacji porzucenia dalszego udziału respondent nie dostanie wynagrodzenia za wykonaną dotychczas pracę.

na dwie, a 19,8% ($n$ = 51) na jedno pytanie KU. Dwóch respondentów (0,8%) przeoczyło wszystkie kontrole sprawdzające uwagę. Obserwacje respondentów $n$ = 53, którzy przeoczyli więcej niż jedną kontrolę uwagi, zostały odrzucone.

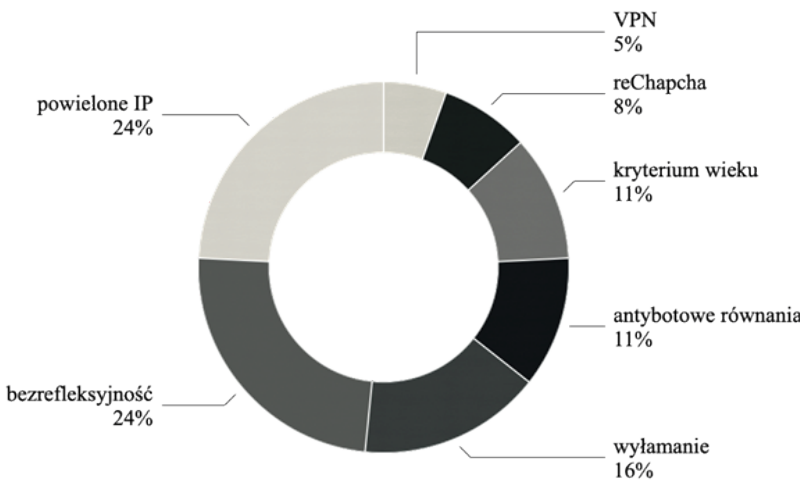
### KONTROLA LOKALIZACJI VPN

Qualtrics umożliwia zapis IP respondenta. Podczas analizy danych wykorzystano formułę makro obliczającą prawdopodobieństwo łączenia się badanego poprzez farmę serwerową. Do kalkulacji użyto narzędzie oferowane przez platformę https://www.ipqualityscore.com. Wykorzystano sugerowaną skalę i stworzono cztery grupy ryzyka nadużycia: minimalne 94,63% ($n$ = 194), niskie 8,29% ($n$ = 17), umiarkowane 17,07% ($n$ = 35) oraz wysokie 5,85% ($n$ = 12). Odrzucono wyniki z grupy wysokiego ryzyka $n$ = 12.

### KONTROLA WIELOKROTNEGO DOSTĘPU DO KWESTIONARIUSZA (*MULTIPLE ENTRY*)

Korzystając z narzędzia Qualtrics, zablokowano możliwość udostępniania linku do badania oraz włączono funkcję blokowania ponownego wejścia w łącze z tego samego adresu IP. Kontrola tego typu miała na celu uniemożliwienie ponownego udziału w badaniu. Podczas analizowania danych zidentyfikowano $n$ = 53 (18,13%) zduplikowanych adresów IP. Powielone numery IP świadczą o bardzo wysokim prawdopodobieństwie wielokrotnego udziału tych samych respondentów w badaniu pomimo blokady Qualtrics. Dane te zostały odrzucone.

**Rysunek 1.** Klasyfikacja źródeł danych niskiej jakości.




Źródło: opracowanie własne

Po wykluczeniu wszystkich danych niskiej jakości $n = 202$ pozostało $n = 140$ wyników. Stanowi to 40,94% całości zebranych obserwacji. Rysunek 1 przedstawia klasyfikację odrzuconych obserwacji.

W końcowej analizie danych zwrócono uwagę na czas wypełniania kwestionariusza. Został on zmierzony i zapisany w sekundach przez Qualtrics. Wynik przeliczono na minuty. Znacząca część obserwacji charakteryzowała się wyraźnie krótkim czasem wypełnienia. W celu operacjonalizacji czasu, jaki należy przyjąć za potrzebny na wypełnienie, kwestionariusz wraz z instrukcją został przeliczony na liczbę słów. Zgodnie z koncepcją opisaną przez Bell (2001) przyjęto, że średni czas przeczytania kwestionariusza zawierającego 1553 wyrazy powinien zająć czytającym szybko 4,43, umiarkowanie – 6,21 a wolno – 10,35 minuty. Założono, iż każde pytanie wymaga dodatkowych 2 sekund na zaistnienie niezbędnych procesów poznawczych oraz zaznaczenie odpowiedzi. Do przyjętych czasów przeczytania dodano 222 sekundy, czyli 3,7 minuty (+2 sekundy dla każdego ze 111 pytań ankiety). Tabela 1 przedstawia estymację czasu wypełnienia kwestionariusza dla badanych czytających w różnym tempie.

**Tabela 1.** Operacjonalizacja czasu oszacowanego na wypełnienie ankiety dla grup czytających szybko, umiarkowanie i wolno (w minutach).

| Prędkość czytania | Czas przeczytania | Procesy: zrozumienie, wyszukanie, osąd i odpowiedź | Czas wypełnienia |
|---|---|---|---|
| Szybka | 4,43 | 3,7 | 8,07 |
| Umiarkowana | 6,21 | 3,7 | 9,91 |
| Wolna | 10,35 | 3,7 | 14,05 |

Źródło: opracowanie własne

Zgodnie z przyjętymi założeniami obliczono, że najkrótszy czas wypełnienia kwestionariusza powinien wynosić około 8 minut. Szybsze tempo odpowiadania może sugerować formę bezrefleksyjności, np. przeklikiwanie. W czasie krótszym niż założone 8,07 minut kwestionariusz wypełniło 46 osób. Stanowi to 32,6% z pozostałych 140 wyników. Zaskakujące jest, że większość osób z krótkimi czasami wypełnienia odpowiedziała bezbłędnie na wszystkie kontrole uwagi. Pięciu najszybszych uczestników badania (od 2 do 5 minut) odpowiedziało bezbłędnie na wszystkie KU. Być może świadczy to o wystąpieniu opisywanego wcześniej zachowania doświadczonych, świadomych i aktywnie poszukujących KU respondentów. Wyłapują oni pytania kontrolne podczas przeklikiwania ankiety. Zaliczając wyniki poniżej 8,07 minut do nieużytecznych, z początkowej puli $n = 342$ jako dane jakościowe rozpoznane zostałoby finalnie jedynie 27,49%, czyli $n = 94$ obserwacje.

Przedstawiony model operacjonalizacji czasu jest teoretycznym konspektem opartym na pewnych założeniach. Z pewnością ludzie różnią się nie tylko biegłością w czytaniu, ale również szybkością procesów przetwarzania informacji. Pozwala on jednak zrozumieć, że ważnym aspektem analizy jakości danych pod kątem bezrefleksyjności respondentów powinna być pewna forma kontroli czasu. Bezrefleksyjność może być rozpoznana pomimo pozytywnych odpowiedzi na pytania typu KU.

## WNIOSKI

Otrzymane wyniki wskazują istotność stosowania odpowiednio dobranych, skutecznych technik pozwalających na weryfikację jakości otrzymywanych wyników w procesie badawczym. Kontrola ponownego wejścia przez użytkownika oferowana przez Qualtrics zdaje się techniką o niewielkiej skuteczności. Rozpoznano wielokrotnie powtarzające się numery IP mogące stanowić istotne źródło zniekształceń zebranych danych. Zawodu dostarczyło również zjawisko licznych prób fałszowania wieku przez respondentów pomimo oferowanej przez ATM możliwości dotarcia do specyficznej próby wiekowej 55+. Dostęp do grup tego rodzaju jest płatny dodatkowo i niestety okazał się wydatkiem o wątpliwej zasadności.

Wyniki kontroli bezrefleksyjności pytaniami KU po konfrontacji z czasem wypełniania wskazują na kolejne wątpliwości. Zdaje się to potwierdzać tezę o powstaniu specyficznej subpopulacji użytkowników biegłych w szybkim wypełnianiu ankiet. Pośpieszne ukończenie ankiety nie sprzyja dostarczaniu relewantnych wyników. Na wyciągnięcie niespodziewanych wniosków pozwoliły dane dostarczone podczas analizy równań antybotowch. Zaledwie dwa pytania pozwoliły wykluczyć aż 25 uczestników, którzy z racji nieznajomości języka nie byli w stanie rozwiązać prostego działania matematycznego. Łatwy test w formie otwartego pytania, którego zadaniem było wyłapywanie działalności algorytmów, pozwolił zidentyfikować relatywnie dużą pulę danych zniekształcających wyniki. Niewystarczająca znajomość języka przez badanych stanowi oczywiste i poważne zagrożenie.

Zastosowane narzędzia i techniki kontroli pozwoliły na identyfikację danych niskiej jakości i wykluczenie 72,5% obserwacji. Pomimo dużego zainteresowania ATM i bardzo szybkiego zebrania wyników wnioski zdają się niejednoznaczne. W odniesieniu do pokazanych statystyk podczas planowania liczby obserwacji należy założyć, iż trzy czwarte danych mogą okazać się bezużyteczne. Korzystanie z ATM skłania do przyjęcia wniosku,

iż badacz planujący zebrać 300 wyników wysokiej jakości powinien zaplanować 1200 uczestników w próbie.

Czytelnik może zastanawiać się nad zasadnością znaczenia badania amerykańskiego panelu typu Mechanical Turk dla rodzimych badaczy. Jednak dynamiczne powstawanie licznych portali dostępnych dla polskich użytkowników może również rodzić trudności podobnej natury. Strony te funkcjonują na bardzo podobnych zasadach co platforma ATM, oferując wynagrodzenie w zamian za wypełnienie ankiety. Trwająca pandemia, niepewność zawodowa i spędzanie znaczącej ilości czasu w izolacji społecznej może sprzyjać chęci udziału w internetowych panelach badawczych, stanowiąc atrakcyjną opcję zarobku. Być może w związku z napływem znaczącej liczby imigrantów ze wschodu można będzie spodziewać się zaistnienia podobnych zjawisk, które obserwujemy na amerykańskim ATM.

Źródła zniekształcenia danych na podstawie literatury przedmiotu oraz badania własnego rodzą pytania o trafność wyników uzyskiwanych w badaniach kwestionariuszowych online przeprowadzonych za pośrednictwem innych paneli. Zasadne i konieczne zdaje się zweryfikowanie użyteczności danych badawczych uzyskiwanych za ich pomocą.

## BIBLIOGRAFIA

Anduiza, E., Galais, C. (2017). Answering Without Reading: IMCs and Strong Satisficing. *Online Surveys International Journal of Public Opinion Research. 29(3).* 497–519. Pobrano z: https://academic.oup.com/ijpor/article-abstract/29/3/497/2669464 (10.01.2021).

Becker, R., Möser, S., Glauser, D. (2019). Cash vs. vouchers vs. gifts in web surveys of a mature panel study – main effects in a long-term incentives experiment across three panel waves. *Social Science Research. 81.* 221–234. Pobrano z: https://www.sciencedirect.com/science/article/pii/S0049089X18305581 (10.01.2021).

Bell, T. (2001). Extensive reading: Speed and comprehension. *The Reading Matrix. 1(1).* Pobrano z: https://www.semanticscholar.org/paper/Extensive-Reading%3A-Speed-and-Comprehension.-Bell/12f10caba81be9aa363fe1d92d4aac97fc025d55?p2df (10.01.2021).

Bosnjak, M., Poggio, T., Becker, K.R., Funke, F., Wachenfeld, A., Fischer, B. (2013). Online survey participation via mobile devices. W: *The American Association for Public Opinion Research (AAPOR) 68th Annual Conference*. Pobrano z: http://www.aapor.org/AAPOR_Main/media/AnnualMeetingProceedings/2013/Session_I-5–2-Bosnjak.pdf (9.01.2021).

Chandler, J., Shapiro, D. (2016). Conducting Clinical Research Using Crowdsourced Convenience Samples. *Annual Review of Clinical Psychology. 12*. 53–81. Pobrano z: https://www.annualreviews.org/doi/full/10.1146/annurev-clinpsy-021815–093623 (9.01.2021).

Everett, A.C.J., Earp, B. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology. 6*. Pobrano z: https://www.frontiersin.org/articles/10.3389/fpsyg.2015.00621/full (9.01.2021).

Hargittai, E., Shaw, A. (2020). Comparing Internet Experiences and Prosociality in Amazon Mechanical Turk and Population-Based Survey Samples. *Socius. 6*. Pobrano z: https://journals.sagepub.com/doi/pdf/10.1177/2378023119889834 (3.01.2021).

Huang, J.L., Curran P.G., Keeney, J., Poposki, E.M., DeShon, R.P. (2012). Detecting and Deterring Insufficient Effort Responding to Surveys. *Journal of Business and Psychology. 27(1)*, 99–114. Pobrano z: https://journals.sagepub.com/doi/pdf/10.1177/2378023119889834 (7.01.2021).

Kees, J., Berry, C., Burton S., Sheehan, K. (2017). An Analysis of Data Quality: Professional Panels, Student Subject Pools, and Amazon's Mechanical Turk. *Journal of Advertising. 46(1).* 141–155. Pobrano z: https://www.tandfonline.com/doi/abs/10.1080/00913367.2016.1269304 (3.01.2021).

Kung, F.Y.H., Kwok, N., Brown, D.J. (2018). Are Attention Check Questions a Threat to Scale Validity? *Applied Psychology: An International Review. 67(2)*, 264–283. Pobrano z: https://iaap-journals.onlinelibrary.wiley.com/doi/full/10.1111/apps.12108 (2.01.2021).

Lovett, M., Bajaba, S., Lovett, M.M., Simmering, M.J. (2018). Data Quality from Crowd sourced Surveys: A Mixed Method Inquiry into Perceptions of Amazon's Mechanical Turk Masters. *Applied Psychology: An International Review. 67(2)*, 339–366. Pobrano z: https://iaap-journals.onlinelibrary.wiley.com/doi/full/10.1111/apps.12124 (14.01.2021).

Mancosu, M., Ladini, R., Vezzoni, C. (2019). 'Short is better'. evaluating the attentiveness of online respondents through screener questions in a real survey environment. *Bulletin of Sociological Methodology/ Bulletin de Méthodologie Sociologique*, *141(1)*, 30–45. Pobrano z: https://journals.sagepub.com/doi/full/10.1177/0759106318812788 (2.01.2021).

Meade, A.W., Craig, S.B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*, 437–455. Pobrano z: https://pdfs.semanticscholar.org/41b7/7840bf309358ecf45b16d00053ed12aea5c0.pdf (2.01.2021).

Munger, K. (2020). *Knowledge Decays: Temporal Validity and Social Science in a Changing World*. Pobrano z: https://files.osf.io/v1/resources/3mnzu/providers/osfstorage/5d6d45b980f9b5001763c4d2?action=download&version=1&direct (dostęp: 2.01.2021).

Moss, A., Litman, L. (2018). After the Bot Scare: Understanding What's Been Happening With Data Collection on MTurk and How to Stop It. *CloudResearch*. Pobrano z: https://www.cloudresearch.com/resources/blog/after-the-bot-scare-understanding-whats-been-happening-with-data-collection-on-mturk-and-how-to-stop-it/ (2.01.2021).

Peyton, K., Huber, G.A., Coppock, A. (2020). *The Generalizability of Online Experiments Conducted During The COVID-19 Pandemic*. Pobrano z: https://osf.io/s45yg/download (2.01.2021).

Shamon, H., Berning, C. (2020). Attention Check Items and Instructions in Online Surveys with Incentivized and Non-Incentivized Samples: Boon or Bane for Data Quality? *Survey Research Methods. 14(1).* 55–77. Pobrano z: https://ojs.ub.uni-konstanz.de/srm/article/view/7374/6874 (2.01.2021).

Tourangeau, R., Rips, L.J., Rasinski, K. (2000). The psychology of survey response. Cambridge: Cambridge University Press.

Tourangeau, R., Sun, H., Yan, T., Maitland, A., Rivero, G., & Williams, D. (2017). Web Surveys by Smartphones and Tablets: Effects on Data Quality. *Social Science Computer Review*, *36(5)*, 542–556. Pobrano z: https://www.researchgate.net/profile/Ting_Yan3/publication/318613541_Web_Surveys_by_Smartphones_and_Tablets_Effects_on_Data_Quality/links/597b6f040f7e9b880281afae/Web-Surveys-by-Smartphones-and-Tablets-Effects-on-Data-Quality.pdf (2.01.2021).