RUCH PRAWNICZY, EKONOMICZNY I SOCJOLOGICZNY
Rok LXXXVI – zeszyt 4 – 2024
https://doi.org/10.14746/rpeis.2024.86.4.14

TOMASZ WANAT[a]

WPŁYW ELIMINACJI ODPOWIEDZI NIEUWAŻNYCH I NIETYPOWYCH RESPONDENTÓW NA REPLIKACJĘ WYNIKÓW BADAŃ W NAUKACH SPOŁECZNYCH

THE IMPACT OF ELIMINATING CARELESS RESPONSES AND OUTLIERS ON THE REPLICATION OF RESEARCH FINDINGS IN SOCIAL SCIENCES

Much of scientific research is difficult or even impossible to replicate or reproduce, a phenomenon known as the replication crisis. One contributing factor to this crisis is the poor quality of the data used in research. This can often be attributed to inattentive or atypical respondents. By eliminating data from these groups, the quality of the research data might improve, potentially increasing the likelihood of successful replication. However, this approach can also contribute to the replication crisis. The methods for detecting and removing inattentive and atypical respondents vary significantly, produce different outcomes, and can be applied in numerous ways – adding another layer of complexity to the replication challenge. The main purpose of the article is to point out the risks inherent in using different methods for detecting inattentive and atypical responses in relation to the replicability of survey results. The article is divided into two parts. The first discusses issues related to the sources of the replication crisis in the social sciences and the potential impact of methods for detecting inattentive responses on research replicability. In the second part, based on a case study of one of the surveys posted on Open Science Framework (OSF), the article demonstrates how subtle yet significant the impact of the methods used to detect and remove inattentive and atypical respondents can be on the success of survey replication. The final section identifies steps to reduce the replication problem associated with the use of methods to detect inattentive and atypical responses.

Keywords: careless responding; replication crisis; questionable research practices

Znaczna część badań naukowych jest trudna lub nawet niemożliwa do replikowania lub odtworzenia, co określane jest mianem kryzysu replikacji. Jednym z czynników przyczyniających się do tego kryzysu jest niska jakość danych wykorzystywanych w badaniach. Często można to przypisać nieuważnym lub nietypowym respondentom. Eliminacja danych z tych grup może poprawić jakość danych badawczych i potencjalnie zwiększyć prawdopodobieństwo udanej replikacji. Eliminacja takich danych może czasami mieć skutek odwrotny. Metody wykrywania i usuwania nie-

[a] Poznań University of Economics and Business, Poland /
Uniwersytet Ekonomiczny w Poznaniu, Polska
tomasz.wanat@ue.poznan.pl, https://orcid.org/0000-0001-9429-5564

RPEiS 86(4), 2024: 251–271. *ISSN (Online) 2543-9170 ISSN (Print) 0035-9629*

uważnych i nietypowych respondentów różnią się znacznie, dlatego też dają różne wyniki i mogą być stosowane na wiele sposobów, dodając kolejny poziom złożoności w kontekście replikacji. Głównym celem artykułu jest wskazanie na zagrożenie tkwiące w posługiwaniu się różnymi metodami wykrywania nieuważnych i nietypowych odpowiedzi dla możliwości odtworzenia wyników badania. Artykuł podzielony jest na dwie części. W pierwszej omówiono zagadnienia związane ze źródłami kryzysu replikacji w naukach społecznych i potencjalnego wpływu metod wykrywania nieuważnych odpowiedzi respondentów na możliwości replikowania badań. W drugiej części, na podstawie analizy przypadku jednego z badań zamieszczonych w systemie Open Science Framework (OSF), pokazano, jak subtelny, a zarazem znaczący może być wpływ zastosowanych metod wykrywania i usuwania nieuważnych i nietypowych respondentów na powodzenie replikacji badań. W końcowej części artykułu wskazano na kroki mające na celu ograniczenie problemu z replikacją związaną z wykorzystaniem metod wykrywania nieuważnych i nietypowych respondentów.

Słowa kluczowe: nieuważne odpowiedzi; kryzys replikacyjności; niewłaściwe praktyki badawcze

I. WPROWADZENIE

Na początku XXI w. spory zamęt w środowisku psychologów (Brzeziński, 2023; Maxwell i in., 2015) wywołały działania podjęte przez kilka grup naukowców, między innymi należących do Open Science Collaboration, których celem było odtworzenie na dużych próbach wyników najbardziej znanych badań (Camerer i in., 2018; Nosek i in., 2022). Okazało się, że bardzo wielu z tych tak fundamentalnych dla tej dyscypliny eksperymentów nie udało się replikować. Zrodziło to obawę, że wiedza z psychologii składa się w dużej mierze z fałszywie pozytywnych wyników (Franco i in., 2016; Nosek i in., 2022). Myliłby się jednak ten, kto odetchnąłby z ulgą, że kwestia ta dotyczy wyłącznie psychologii. Problem z replikacją wyników badań rozlał się z siłą tsunami po innych dyscyplinach naukowych, takich jak zarządzanie (Pagell, 2021), ekonomia (Camerer i in., 2018), nauki medyczne (Coiera i in., 2018), socjologia (Freese i Peterson, 2017) czy nauki humanistyczne (Peels i Bouter, 2018), prowadząc do upowszechnienia się w naukach społecznych pojęcia „kryzysu replikacji".

Wyprowadzenie nauki z kryzysu replikacji jest o tyle trudne, że problemy z nim związane wypływają z wielu, niezależnych od siebie, źródeł (Ellis, 2022). Położenie tamy tylko niektórym ze źródeł kryzysu nie jest w stanie uchronić nauki przed zalewem niskiej jakości badań. Co więcej, istnieją działania, które w swoim zamyśle mają poprawiać jakość badań, a tym samym ograniczać możliwości niepowodzenia w przypadku replikacji, a które użyte w niewłaściwy sposób mogą prowadzić do potęgowania tego problemu. Takim działaniem jest między innymi eliminowanie wybranych danych (odpowiedzi od respondentów) uznanych za tzw. nietypowe lub nieuważne (Bakker i in., 2021).

Generalnie przyjmuje się, że dla zapewnienia odpowiedniej jakości danych konieczne jest poddanie ich procesowi dokładnej kontroli (Baillie i in., 2022). Dotyczy to wielu aspektów inspekcji danych, w tym, przykładowo, wykrywania wartości odstających (Leys i in., 2019), a także wykrywania odpowiedzi pochodzących od niezaangażowanych respondentów (Curran, 2016) lub botów, szczególnie groźnych w coraz popularniejszych badaniach online (Bybee i in., 2022;

Ward i Meade, 2023). W literaturze wskazuje się wręcz na konieczność posługiwania się metodami wykrywania nietypowych i nieuważnych respondentów (Ward i Meade, 2023). O ile nie można kwestionować słuszności samej idei dążenia do dobrej jakości danych, o tyle z takimi praktykami związane są pewne problemy. Jeden z nich, wynikający z wielości metod wykrywania nietypowych i niezaangażowanych respondentów, polega na oznakowywania różnych grup respondentów (DeSimone i Harms, 2018). Skutkiem tego pojawia się możliwość selektywnego usuwania wybranych danych, co może powodować trudność z replikowaniem badania przez innych badaczy, zarówno wtedy, gdy metody te nie będą użyte, jak i wtedy, gdy zostaną zastosowane.

Celem tego opracowania jest przedstawienie problemu powstającego na styku replikacji oraz metod związanych z eliminacją odpowiedzi nieuważnych i nietypowych respondentów (NNR). Do lepszego zobrazowania tego problemu wykorzystano analizę przypadku bazującą na danych pochodzących z badania zamieszczonego na platformie Open Science Framework (https://osf.io/dashboard). Dodatkowo wskazano konieczne działania mające na celu, jeżeli nie wyeliminowanie, to chociaż ograniczenie problemów z replikacją wypływających z tego źródła.

Struktura opracowania jest dwuczęściowa. W części pierwszej (punkty I–V) zaprezentowano zagadnienia związane z kryzysem replikacji w naukach społecznych oraz metodami kontrolowania uwagi respondentów i ich wpływu na możliwości replikowania badań. W części drugiej (punkty VI–VII) przedstawiono przykład odtworzenia badania naświetlający złożoność omawianych problemów.

II. ISTOTA REPLIKACJI BADAŃ

Jednym z problemów związanych z replikacją jest fakt niedoskonałego rozumienia samej istoty replikacji. Jak zauważają Nosek i Errington (2020, s. 2), często za replikację uważa się powtórzenie procedury badania i obserwowanie, czy uprzednio uzyskane wyniki badania się powtarzają. Takie podejście pomimo swojej urzekającej prostoty jest błędne z uwagi na zmienność i specyficzność warunków, w jakich przeprowadza się badania. Zamiast przytoczonego wcześniej ujęcia Nosek i Errington (2020) proponują następującą definicję: replikacja to badanie, w którym każdy wynik można uznać za dowód diagnostyczny dotyczący twierdzenia z wcześniejszych badań. W takim ujęciu zakończona powodzeniem replikacja osiąga wyższy poziom generalizacji z powodu odmienności warunków, w jakich oba badania były przeprowadzane. Z kolei nieudana replikacja – lub lepiej w liczbie mnogiej replikacje – wskazuje na ograniczoną wiarygodność wcześniej uzyskanych wyników. Nawet jednak negatywny wynik lub wyniki są o tyle ważne, że dają możliwość stworzenia lepszych koncepcji przyszłych badań. Ogólny mechanizm działania replikacji polega więc na zwiększaniu zaufania do twierdzeń w przypadku zgodnych wyników i obniżaniu go w przypadku niezgodnych. Tego typu symetria promuje

replikację jako mechanizm konfrontacji wcześniejszych twierdzeń z nowymi dowodami (Nosek i in., 2022).

Pojęcie replikacji warto odróżnić od pojęcia odtwarzalność (*reproducibility*). To ostatnie odnosi się do całkowitego powtórzenia eksperymentu na podstawie pełnych danych dotyczących projektu badawczego, informacji i decyzji podjętych podczas kodowania i analizy danych (Aguinis i in., 2018).

III. ŹRÓDŁA KRYZYSU REPLIKACJI

Potencjalnych źródeł kryzysu replikacji jest wiele i zaliczyć do nich można między innymi: problem ze rozumieniem statystycznej istotności (Białek i Wolski, 2023), skłonność do publikowania pozytywnych wyników badań, ograniczony zakres badań replikacyjnych i metaanaliz, występowania tzw. niewłaściwych praktyk badawczych, powszechność badań o niskiej mocy i małej liczebności próby, błędne stosowanie testów statystycznych czy niepełne raportowanie badań (Camerer i in., 2018). Warto się przyjrzeć tym problemom z większą uwagą.

Za jedno z głównych źródeł kryzysu replikacji uznaje się tzw. tendencyjność publikowania (*publication bias*), czyli skłonność do publikowania prac z istotnymi statystycznie wynikami, a niechętne traktowanie przez samych autorów, a także recenzentów oraz wydawców prac z wynikami nieistotnymi statystycznie (Muradchanian i in., 2023). Na fakt, że tak się dzieje, wskazują choćby statystyki pokazujące, że 96% typowych publikowanych badań w psychologii kończy się istotnymi statystycznie wynikami, podczas gdy takich samych pozytywnych rezultatów jest tylko 44% w przypadku wcześniejszej rejestracji programu badawczego (Scheel i in., 2021). Równie wyraźnie tendencyjność publikacji uwidacznia się w metaanalizach (Świątkowski i Dompnier, 2017).

Innym źródłem kryzysu replikacji jest, co brzmi zaskakująco, brak badań replikacyjnych. Te ostatnie mają kluczowe znaczenie dla walidacji i potwierdzenia istniejących ustaleń (Tincani i Travers, 2019). Niestety wydawcy prestiżowych czasopism często traktują priorytetowo publikowanie badań posiadających nowatorski charakter, a pomijają te o charakterze replikacji (Duvendack i in., 2017). Przykładowo, spośród 333 czasopism ekonomicznych tylko 18 opublikowało więcej niż 3 badania będące replikacjami (Duvendack i in., 2017). Bardzo mało jest czasopism systemowo nastawionych na publikowanie badań o charakterze replikacji. Pośrednio do braku badań replikacyjnych przyczynia się ograniczona transparentność publikowanych badań (Pagell, 2021). Jeżeli nie ma szczegółowych informacji o każdym z etapów procedury badawczej, tj. teorii, projektowania, pomiaru, analizy i raportowania wyników, to trudno takie badanie odtworzyć (Aguinis i in., 2018). Tylko nieliczne czasopisma (choć ich liczba rośnie) decydują się na publikowanie danych oraz kodów potrzebnych do odtworzenia badania (Duvendack i in., 2017).

Bardzo „wydajnym" źródłem zasilania kryzysu replikacji jest stosowanie przez badaczy tzw. niewłaściwych praktyk badawczych (*questionable research*

practices [QRP]; Bakker i in., 2021; Ellis, 2022), odnoszących się do działań, które, choć nie zawsze jawnie nieetyczne, mogą poważnie zniekształcać wyniki przeprowadzanych analiz. Te praktyki obejmują szeroki zakres działań, przy czym samo wymienienie zestawu takich praktyk zajęłoby zbyt wiele miejsca, jako że w literaturze przywołuje się kilkadziesiąt ich typów (Wicherts i in., 2016). W tym miejscu warto wskazać kilka najbardziej typowych, począwszy od „subtelnych", jak *p-hacking* (czyli manipulowanie analizą danych w celu osiągnięcia istotnie statystycznych wyników), poprzez bardziej bezpośrednie formy manipulacji jak HARKing (czyli stawianie hipotez po poznaniu wyników; Kerr, 1998), po jednoznacznie nieetyczne praktyki jak fałszowanie danych i wyników.

Jednym z najbardziej typowych działań w zakresie p-hackingu jest wielokrotne testowanie hipotez na tej samej bazie danych dla różnych hipotez bez korekty na wielokrotne porównania. Gdy przeprowadza się wiele testów statystycznych na tej samej bazie danych, prawdopodobieństwo popełnienia przynajmniej jednego błędu typu I rośnie z każdym dodatkowym testem. Dla przykładu, przeprowadzenie 20 niezależnych testów przy $\alpha = 0{,}05$ daje już ponad 64% szans na wystąpienie co najmniej jednego fałszywie pozytywnego odkrycia.

Inną popularną formą p-hackingu jest nieortodoksyjne podejście do wyboru zmiennych. Badacze mogą wypróbować wiele analiz z różnymi kombinacjami zmiennych niezależnych, zależnych, moderujących, kowariancyjnych bądź mediujących. Na przykład analiza regresji zawierająca 20 niezależnych zmiennych binarnych da więcej niż 1 milion różnych wyników przez uwzględnienie każdego poziomu każdej zmiennej we wszystkich możliwych kombinacjach.

Kolejnym przykładem działań związanych z p-hakingiem jest uzupełnianie brakujących danych bez zgłaszania, że dane te zostały imputowane. Relatywnie często w badaniach napotyka się problem braku danych. Rozwiązaniem tego problemu może być albo usunięcie danych respondentów z brakującymi danymi, co obniża wielkość badanej próby i dalej moc testów, albo uzupełnia dane poprzez imputację (Wicherts i in., 2016). Ta polega na zastępowaniu brakujących danych estymowanymi wartościami. Problemem nie jest w tym przypadku sam fakt uzupełniania danych, ale brak informacji o tym fakcie. Może to mieć poważne konsekwencje dla interpretacji wyników, w szczególności gdy brak danych nie jest losowy (Yang i in., 2019). W kontekście replikacji imputowanie danych jest trudne do odtworzenia dlatego, że zbiór respondentów nieudzielających odpowiedzi nie musi składać się z osób o tych samych charakterystykach.

Wskazana powyżej elastyczność p-hackingu w osiąganiu istotnie statystycznych wyników sprawia, że przyznaje się do niej blisko 63% badaczy z USA, choć już „tylko" 47% z Włoch (Agnoli i in., 2017).

Inną formą niewłaściwych praktyk badawczy jest tzw. HARKing (*Hypothesizing after the Results are Known*; Andrade, 2021). Polega on na formułowaniu lub modyfikowaniu hipotez badawczych po zapoznaniu się z wynikami eksperymentu lub analizy, a nie przed ich przeprowadzeniem (Bridges, 2022). Niemniej HARKing jest o tyle problematycznym działaniem, że two-

rzy fałszywe wrażenie zaprojektowania badania w celu testowania określonej hipotezy, podczas gdy w rzeczywistości hipoteza została dostosowana do już uzyskanych wyników. Szacuje się, że zarażonych HARKingiem może być w niektórych dyscyplinach nawet 50% badań (Banks i in., 2016), może on też przybierać różne formy, począwszy od sformułowania hipotezy, poprzez uzyskanie wyników, po tworzenie niejasnej hipotezy, która na przykład nie ma określonego kierunku zależności pomiędzy zmiennymi (np. „Zwycięstwo Donalda Trumpa w wyborach prezydenckich w USA wpłynie na ceny akcji na giełdzie" – przy czym nie wiadomo, czy wpłynie na wzrost czy spadek cen akcji), lub też dzielenie badanej populacji na grupy tak, aby hipoteza potwierdziła się chociaż w jednej z nich.

Jeszcze inny źródłem problemów z replikacją jest niska moc statystyczna przeprowadzanych analiz (Brzeziński, 2023). Moc statystyczna to prawdopodobieństwo prawidłowego odrzucenia hipotezy zerowej przy założeniu, że hipoteza alternatywna jest prawdziwa (Lakens i Evers, 2014). Badania o zbyt małej mocy mają niskie prawdopodobieństwo wykrycia prawdziwego efektu, a w takiej sytuacji nie są publikowane ze względu na wspomnianą wcześniej tendencyjność publikowania. Niemniej jeśliby nawet badanie o małej mocy wykryło jakiś efekt, to jego moc predykcyjna będzie niska i często zawyżona (Ioannidis, 2008). Problem ten dość wyraźnie pokazują przeprowadzane metaanalizy. Jedno z takich badań wskazało, że średnia wielkość efektu w badaniach społeczno-psychologicznych w okresie ostatnich 100 lat wynosi zaledwie $r = 0{,}21$ (Richard i in., 2003).

Kolejnym źródłem problemów z replikacją jest błędne zastosowanie testów statystycznych i metod statystycznych (Vowels, 2023). W najprostszym wariancie może ono dotyczyć użycia jednostronnego testu t (choćby w celu łatwiejszego osiągnięcia poziomu istotności), gdy bardziej uzasadniony jest dwustronny test t. Badanie replikacyjne, o ile nie zastosuje ponownie nieuzasadnionego koncepcją badania testu jednostronnego, może dać wyniki nieistotne. Dość typowym przykładem błędnego stosowania testów statystycznych jest sytuacja prezentowania wyników, dla których nie są spełnione niezbędne założenia na przykład homogeniczności wariancji czy normalności rozkładu zmiennych (Stodden, 2015), typowe dla wielu standardowo wykorzystywanych metod analizy danych jak ANOVA czy SEM. Do tej grupy problemów jako potencjalne źródło trudności z replikacją zaliczyć też można brak walidacji konstruktów wykorzystywanych w badaniach (Leichtmann i in., 2022).

Dodatkowym problemem związanym z replikacja badań jest aspekt komunikacyjny dotyczący standardów raportowania badań. Związany jest z kompletnością i precyzją, z jaką opisuje się plan badań, jego przebieg, analizy i wnioski statystyczne (Leichtmann i in., 2022). W celu umożliwienia jak najwierniejszej replikacji wskazane jest raportowanie wszystkich zastosowanych metod i procedur badawczych oraz wyników analiz, łącznie z tymi, które prowadziły do nieistotnych rezultatów. Równie ważne jest nie tylko pełne raportowanie statystyk, w tym stopni swobody i statystyk testowych, lecz także wielkości efektów i przedziałów ufności, a nie tylko wartości p (Lakens i Evers, 2014).

Tak jak liczne są przyczyny kryzysu replikacji, tak również liczne są propozycje mające na celu wyeliminowaniu lub choćby ograniczenie tego problemu (Brzeziński, 2023). Warto w tym miejscu choć pokrótce wspomnieć o szeregu z nich, takich jak: rejestrowanie planów badawczych i hipotez przed rozpoczęciem badań (*pre-registration*); ustanowienie standardów czasopism dotyczących publikacji badań z nieistotnymi wynikami (Tincani i Travers, 2019), zachęcanie do prowadzenia i publikowania replikacji badań, zarówno bezpośrednich (powtórzenie tej samej metody), jak i konceptualnych (testowanie tych samych hipotez za pomocą innych metod); zwiększenie roli metaanaliz i systematycznych przeglądów literatury wspomagających identyfikację trendów i wzorców w danych, co może ujawnić potencjalne problemy z replikacją; zapewnienie pełnej transparentności materiału badawczego przez udostępnianie surowych danych, materiałów badawczych oraz instrukcji pozwala innym naukowcom na dokładniejsze zrozumienie i powtórzenie badań (Aguinis i in., 2018); zachęcanie do publikowania wyników, które nie potwierdzają hipotez badawczych (Franco i in., 2016), ale mają odpowiednio dużą moc, co może zapobiegać występowaniu opisywanej wcześniej tendencyjności publikowania; elastyczne podejście do wartości *p*, przez co należy rozumieć położenie większego nacisku na wskaźniki wielkości efektu i na przedziały ufności (Brzeziński, 2023).

IV. METODY WYKRYWANIA NIEUWAŻNYCH I NIETYPOWYCH RESPONDENTÓW

Metody i techniki wykrywania nieuważnych i nietypowych respondentów (NNR) mają kluczowe znaczenie w kontekście kryzysu replikacyjnego w naukach społecznych. Jakość danych uzyskiwanych z badań ankietowych może być znacząco obniżona przez uczestników, którzy nie odpowiadają na pytania w sposób przemyślany i rzetelny, ale na przykład stosują strategię satysfakcjonowania ograniczającą wysiłek wkładany w udzielanie odpowiedzi (Krosnick, 1991). Nieuważni i nietypowi respondenci mogą prowadzić do fałszywie pozytywnych lub negatywnych wyników, co wpływa na wiarygodność badań i ogranicza możliwość ich replikacji. Jednakże niewłaściwe stosowanie tych metod może paradoksalnie potęgować kryzys replikacyjności. Na przykład nadmierne oczyszczanie danych z odpowiedzi uznanych za nieuważne, bez solidnych podstaw teoretycznych i metodologicznych, może prowadzić do eliminacji ważnych informacji, co zniekształca wyniki i utrudnia replikację.

Metod i technik wykrywających NNR jest co najmniej kilkadziesiąt, nie licząc ich szczegółowych podtypów. Po części wynika to z faktu, że każda z nich jest w stanie wykryć tylko pewien rodzaj nieuważności lub nietypowości. Dobrych przeglądów tych metod i technik dostarczają publikacje Curran (2016), Huang i in. (2012) oraz Ward i Meade (2023), a w polskiej literaturze np. Saad (2021).

Metody i techniki wykrywania NNR skupiają się na detekcji pewnych wzorców odpowiedzi lub zachowań, które mogą świadczyć o nieuważnym udzielaniu odpowiedzi na pytania. Przykładowo, osoby niezaangażowane zaznaczają

często te same odpowiedzi na skali. Możliwe to będzie do wykrycia metodą długich ciągów, w której mierzy się maksymalną liczbę tych samych odpowiedzi (Curran, 2016). Osoba, która zaznaczałaby w ankiecie tę samą wartość na skali (np. prawie za każdym razem 7), mogłaby być uznana za nierzetelną. Takie zachowanie możliwe też będzie do wykrycia metodą IRV (Dunn i in., 2018) z uwagi na niski poziom wariancji odpowiedzi lub analizę czasu odpowiedzi, która pozwala zidentyfikować osoby odpowiadające zbyt szybko, by przemyśleć swoje odpowiedzi.

Można też posługiwać się metodami, które sprawdzają, czy respondenci czytają pytania. Wykorzystać do tego można metodę samoopisu, a więc pytanie wprost o poziom uwagi przy wypełnianiu ankiety, można zastosować pytania podchwytliwe, w przypadku których tylko jedna odpowiedź jest prawdziwa (np. *Jestem prezydentem Ukrainy*; Meade i Craig, 2012) lub pytania instruujące, które nakazują, jaką odpowiedź respondent ma zaznaczyć (Oppenheimer i in., 2009) – przykładowo w pytaniu *Jakie znasz marki piwa?* w drugiej jego części można zawrzeć polecenie *To jest test uwagi, zaznacz odpowiedź 1 – nie znam żadnych marek*. Jeżeli ankiety wypełniane są online, można wykorzystać urządzenia techniczne lub specjalne oprogramowanie do analizy ruchów myszą czy pytania typu reCAPTCHA (Saad, 2021). Można też badać spójność odpowiedzi za pomocą analizy synonimów i antonimów psychograficznych bądź metody Mahalanobisa, która wyznacza, w jakim zakresie w przestrzeni wielowymiarowej odpowiedzi danego respondenta różnią się od średniej dla całej populacji badanej (Curran, 2016).

Przytoczony powyżej pobieżny przegląd metod wykrywania NNR wskazuje, że liczba i różnorodność tych metod już sama w sobie jest problemem. Replikacja powinna bowiem zastosować ten sam lub zbliżony zestaw metod co badanie oryginalne. Nie jest to jednak jedyny problem. Drugi związany jest z faktem niewielkiego poziomu substytucyjności pomiędzy poszczególnymi metodami. Dodatkowo dochodzi do tego problem określenie progu rozdzielającego uważnych od nieuważnych respondentów, który w przypadku wielu metod ma do pewnego stopnia uznaniowy charakter.

V. NEGATYWNY WPŁYW METOD WYKRYWANIA NNR NA REPLIKACJĘ BADAŃ

W literaturze naukowej, zarówno tej poświęconej wykrywaniu nieuważnych respondentów, jak i tej związanej z kryzysem replikacyjności, nie wspomina się o możliwości negatywnego wpływu metod wykrywania nieuważnych respondentów na replikacyjność badań. Wskazuje się na bardziej ogólny efekt, jakim jest manipulowanie zbiorem danych (Wicherts i in., 2016). Nie ulega więc wątpliwości, że wpływ taki zachodzi, co więcej, może ujawniać się w wielu miejscach. Aby dokładnie zrozumieć, jak często i w jakim stopniu to zagrożenie wpływa na badania replikacyjne, warto przyjrzeć się temu problemowi z większą wnikliwością.

Przede wszystkim warto zwrócić uwagę na niekomparatywność prób badawczych, w których wykorzystano metody testowania NNR w porównaniu z tymi, w których takich metod nie zastosowano. Abstrahując od kwestii, które podejście jest metodycznie bardziej poprawne w danej sytuacji badawczej, łatwo zauważyć, że struktura populacji badanej w obu przypadkach będzie inna, co w oczywisty sposób może wpłynąć na uzyskiwane wyniki. Dotychczasowe badania wskazują, że nawet niewielka zmiana struktury badanej próby może wpływać na korelacje pomiędzy pozycjami skal (Credé, 2010), oszacowania dotyczące rzetelności skal (Maniaci i Rogge, 2014), wysokość ładunków czynnikowych (Kam i Meyer, 2015), czy też wskazywać na istnienie większej liczby wymiarów, niż to jest w rzeczywistości (Cornell i in., 2012). Od braku komparatywności prób badawczych do problemów z replikacją nie jest więc daleko.

Na tej samej zasadzie działa też niekomparatywność prób badawczych wykorzystujących różne zestawy metod wykrywania NNR. Przykładowo, zastosowanie metryki Mahalanobisa D, a więc metody wykrywającej wartości odstające, będzie miało zupełnie inny skutek niż zastosowanie metryki długich ciągów (Longstrings) wykrywającej ciągi bardzo podobnych do siebie odpowiedzi (które najczęściej są „bardzo typowe"). W obu przypadkach zostaną usunięte grupy respondentów, choć zasadniczo będą to dwie różne grupy. Wyniki mogą być również niekomparatywne w sytuacji zastosowania tej samej metody wykrywania NNR, ale przy przyjęciu innego progu nieuważności lub nietypowości.

Innym przykładem negatywnego wpływu metod wykrywania NNR na replikacyjność badań jest problem z nadmierną eliminacją danych. Metod wykrywania NNR jest wiele i są w stanie oznakować czasami znaczący odsetek badanych respondentów (DeSimone i Harms, 2018). W literaturze wskazuje się na przypadki, gdy wykorzystanie wielu metod wykrywania NNR prowadzi do oznakowania nawet 75% badanych (Vecchio i in., 2020). Przy tak dużym ograniczeniu liczebności próby moc testów statystycznych znacząco spada, a szansa na uzyskanie przypadkowych, ale istotnych statystycznie wyników wzrasta. W konsekwencji stosowanie metod wykrywania NNR może, pomimo dobrych intencji, prowadzić do publikowania większej liczby przypadkowych wyników.

Warto też zwrócić uwagę na kwestie związane z różnicami międzykulturowymi i językowymi w zakresie stylu odpowiadania (He i in., 2014) i ich wpływem na wykrywanie NNR. Ta sama metoda użyta w różnych krajach może wykryć różną liczebność respondentów tzw. nieuważnych ze względu na typowe dla danej kultury style odpowiedzi. W niektórych kulturach istnieje większa skłonność do zaznaczania środkowych pozycji na skali, a w innych (indywidualistycznych) bardziej skrajnych pozycji. Posłużenie się więc tą samą metodą i procedurą wykrywania NNR będzie skutkowało odrzucaniem innych typów respondentów. W istocie także w przypadku metod wykrywania nieuważności konieczne jest, analogiczne do przypadku skal, dokonanie kulturowej adaptacji metod.

O ile przedstawione powyżej przykłady negatywnego wpływu metod wykrywania nieuważnych respondentów na replikacyjność badań nie miały charakteru celowego, a raczej były ubocznym skutkiem stosowania tych metod, o tyle pewne działania podejmowane przez badacza mogą w sposób świadomy i celowy prowadzić do zniekształceń skutkujących mniejszym prawdopodo-

bieństwem replikacji. Chodzi o działania związane z p-hackingiem lub nierzetelnym raportowaniem.

Biorąc pod uwagę fakt, że za pomocą zbioru metod wykrywania nietypowych i nieuważnych respondentów można oznakować różne grupy, dość łatwo można to wykorzystać do uzyskania istotnych statystycznie wyników. Wyobraźmy sobie, że badacz nie uzyskał istotnych statystycznie wyników. Może w takiej sytuacji wyrzucić badanie do kosza lub „przekonać" samego siebie, że brak istotności wynika z niedoskonałości próby badawczej. Nawet jeżeli przed badaniem badacz nie zakładał stosowania żadnej metod wykrywania NNR, to zawsze może skorzystać z tych, które są dostępne już po uzyskaniu wyników. Przyjmijmy, że badacz zastosuje pięć takich metod (przykładowo, badanie długich ciągów odpowiedzi, IRV, odległość Mahalanobisa, korelację Ogół-Osoba, metodę odd-even). Dla trzech z nich (tj. długich ciągów, korelacji Ogół-Osoba, IRV) może dodatkowo ustalić progi na trzech różnych poziomach. W praktyce otrzymuje więc 11 różnych grup respondentów, które może usunąć. Ma więc jeszcze 11 dodatkowych szans na „znalezienie" istotności w ramach pozyskanych danych. Dodatkowo dochodzą do tego kombinacje wspomnianych powyżej metod, więc liczba opcji jest w praktyce jeszcze większa. Tego typu działanie z całą pewnością można zaliczyć do niewłaściwych praktyk badawczych.

Drugą wątpliwą etycznie praktyką jest niepełne raportowanie o wykorzystanych w badaniach metodach wykrywania NNR i procedur postępowania po ich wykryciu. Brak raportowania jest oczywiście standardowym źródłem kryzysu replikacji (Cockburn i in., 2020). Brak raportowania o metodach wykrywania NNR jest tylko jednym z jego wariantów. Niepełne raportowanie jest o tyle poważnym problemem, że dotyczy wielu, czasami szczegółowych, informacji dotyczących zarówno metod, jak i progów, a także kolejności zastosowania metod wykrywania NNR i sposobu postępowania z nimi. Bardzo rzadko, nawet w prestiżowych czasopismach, można spotkać się ze szczegółowymi opisami użytych metod i uzasadnieniem ich użycia. Z problemem raportowania związany jest inny, dotyczący braku standaryzacji wykorzystanych metod wykrywania NNR. Sugestie płynące z literatury (Ward i Meade, 2023) mają ogólny charakter, a na dodatek nie są jeszcze powszechnie znane i akceptowane.

VI. ANALIZA PRZYPADKU BADANIA Z WYKORZYSTANIEM METOD WYKRYWANIA NNR

W celu zilustrowania wagi i złożoności problemu związanego z wpływem wykorzystania metod wykrywania NNR na replikacyjność badań przedstawiona zostanie analiza przypadku bazująca na artykule „Examining the link between social exclusion and social-risk taking: A correlational and experimental investigation" autorstwa dwóch niemieckich badaczek Michael Pfundmair oraz Evy Lermer, będącego w stadium recenzji przed ostateczną publikacją. Artykuł oraz dane dostępne są w systemie OSF (https://osf.io/umrht).

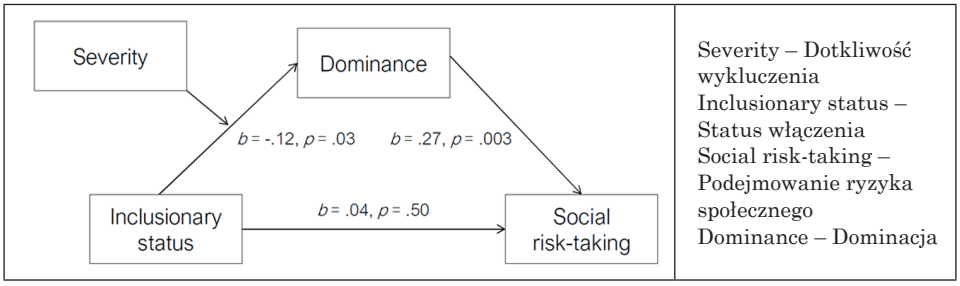
Wykorzystany do odtworzenia eksperyment został wybrany nie ze względu na jego słabości, ale wręcz przeciwnie, z uwagi na staranność przygotowania, przemyślany sposób realizacji, poprawność metodyczną badania oraz właściwe raportowanie zgodne z zasadami *open science*, czyli wstępne zarejestrowanie procedur i założeń badawczych oraz udostępnienie danych.

1. Opis oryginalnego badania

Badanie miało na celu wykazanie, że wykluczenie społeczne zmniejsza skłonność do ryzyka społecznego, w przeciwieństwie do ryzyka związanego z innymi obszarami. Jednym z zamierzeń badania było wykazanie interakcji pomiędzy zmiennymi (zob. rysunek 1). Zakładano, że wykluczeni uczestnicy będą doświadczali niższego poziomu dominacji niż włączeni uczestnicy w łagodnym zadaniu Cyberball[1], ale nie w ekstremalnym zadaniu Future-Life[2], co przełoży się na niższą skłonność do ryzyka społecznego. Jak można zauważyć na podstawie danych zamieszczonych na rysunku 1, założenie to zostało spełnione.

Rysunek 1

Moderowana mediacja w badaniu Pfundmair i Lermer



Źródło: opracowanie własne na podstawie danych z badania Pfundmair i Lermer (2023).

Model moderowanej mediacji był jednym z trzech testowanych modeli, w których analizowano oddzielnie różne zmienne mediujące (oprócz *dominacji* były to *przyjemność* i *pobudzenie*). Tylko w przypadku *dominacji* uzyskano istotne statystycznie rezultaty. W dwóch pozostałych przypadkach nie uzyskano efektu moderowanej mediacji.

[1] W grze komputerowej Cyberball uczestnicy są losowo przydzieleni do jednego z dwóch warunków eksperymentalnych: wykluczenia społecznego lub włączenia społecznego. Uczestnicy badania muszą podawać piłkę dwóm innym wirtualnym graczom. Osoby wykluczone otrzymują bardzo rzadko podania od wirtualnych graczy. Głównie obserwują, jak pozostali uczestnicy gry wymieniają się podaniami, co wywołuje wrażenie wykluczenia.

[2] W zadaniu Future-Life badani są poproszeni o wypełnienie testu osobowości, a następnie otrzymują (fikcyjne) informacje zwrotne na temat ich przyszłego życia.

Badanie miało charakter eksperymentu laboratoryjnego. Jest to o tyle ważne, że fizyczna obecność osób eliminuje niektóre źródła nieuważności (np. wielokrotne odpowiedzi tych samych osób, boty, osoby nieznające lub słabo znające język, w jakim kwestionariusz jest napisany).

W oryginalnym badaniu zastosowano jedną metodę kontrolowania uwagi respondentów. Była to kontrola manipulacji eksperymentalnej odnosząca się do jednej ze zmiennych. Autorki nie wyszczególniają innych metod wykrywania nietypowych i nieuważnych respondentów.

2. Replikacja badania

Odtworzenie badania z wykorzystaniem metod wykrywania NNR przeprowadzono w trzech etapach:

1. Wyznaczenie modelu odtworzonego w celu określenia stopnia porównywalności wyników pomiędzy wynikami uzyskanymi przez Pfundmair i Lermer (2023) a tymi uzyskanymi na potrzeby badania NNR.

2. Oznaczenie respondentów jako nieuważnych lub nietypowych za pomocą czterech wybranych metod.

3. Określenie poziomu istotności zależności w modelu moderowanej mediacji w zależności od użytych metody wykrywania NNR.

Etap 1. Odtworzenie modelu oryginalnego

W pierwszym etapie spróbowano odtworzyć wyniki uzyskane w oryginalnym badaniu. W tym celu posłużono się tym samym oprogramowaniem co w oryginalnym badaniu oraz tymi samymi danymi. Z uwagi na brak pełnych informacji o zastosowanej procedurze istnieje możliwość niedoskonałego odtworzenia procedury badania, a tym samym wyników. W badaniu odtworzonym wykorzystano następującą składnię dla dodatku PROCESS ver. 4.2.dla SPSS v.28 stworzonego przez Hayesa:

```
Process y=Risk_s/x=Inclu/m=Dom_post/w=Severty/cov=Dom_pre /model=7 /center=2 /
                intprobe=10 /conf=95 /jn=1 /boot=5000 / seed=12345,
```

gdzie poszczególne zmienne są oznaczane jako:

– Inclu – Inclusory status – Status społecznego włączenia
– Risk_s – Social risk taking – Podejmowanie ryzyka społecznego
– Dom_post – Dominacja – Pomiar po manipulacji eksperymentalnej
– Dom_pre – Dominacja – Pomiar przed manipulacją eksperymentalną
– Severty – Severity – Dotkliwość wykluczenia.

Uzyskane wyniki, w części dotyczącej zależności o moderującym i mediującym charakterze, przedstawiono w tabeli 1. Obok oryginalnych wyników z publikacji Pfundmair i Lermer (2023) zamieszczono wyniki odtworzone. Odtworzone wyniki w zasadzie nie różniły się od oryginalnych (nie licząc kwestii zaokrąglenia wyniku i liczby stopni swobody[3]).

[3] 167 stopni swobody można uzyskać bez uwzględnienia zmiennej kowariancyjnej Dominance_pre, ale wtedy efekt interakcji staje się nieistotny statystycznie.

Tabela 1

Wyniki analizy moderowanej mediacji: oryginalne i odtworzone

Ścieżka	Model	*b*	*SE*	*t (df)*	*p*
Inclusion → Dominance	Oryginalny	0,16	0,05	3,07 (167)	0,003
	Odtworzony	0,1649	0,0537	3,0734 (166)	0,0025
Severity → Dominance	Oryginalny	0,12	0,05	2,16 (167)	0,03
	Odtworzony	0,1161	0,0538	2,1602 (166)	0,0322
Severity*Inlcusion → Dominace	Oryginalny	–0,12	0,05	–2,15 (167)	0,03
	Odtworzony	–0,1152	0,0535	–2,1535 (166)	0,0327

Źródło: Pfundmair i Lermer (2023) oraz wyniki badania odtworzonego przez autora.

Podobieństwo wyników modelu oryginalnego oraz odtworzonego w przypadku zastosowania procedury bootstrapingu nie zawsze jest pewne. W tym przypadku wyniki są na tyle do siebie zbliżone, że można założyć, iż udało się z powodzeniem odtworzyć model z publikacji Pfundmair i Lermer (2023). Można więc uznać, że kolejne analizy przebiegałyby w podobny sposób w przypadku użycia metod wykrywania NNR, gdyby te zostały przez autorki zastosowane. Dla większej porównywalności wyników bootstrapingu użyto w składni tych samych wartości dla polecenia seed.

Etap 2. Opis wykorzystanych metod wykrywania NNR

Z uwagi na fakt, że w oryginalnym badaniu nie wykorzystano (poza jedną) metod wykrywania nieuważnych respondentów możliwe było wykorzystanie tylko metod *ex post*. Wybrano dwie metody związane z wykrywaniem braku uwagi i niskiego zaangażowania: metodę długich ciągów (*longstrings*) oraz metodę wykorzystującą spójność odpowiedzi na pozycje odwrócone na skali, tzn. MAD (*mean absolute difference*). Natomiast w odniesieniu do wartości nietypowych wybrano odległość Mahalanobisa liczonego dla skal oraz metodę badania korelacji Ogół-Osoba (Total-Person correlation). Sposób obliczenia poszczególnych metryk przedstawia się następująco:

Mahalanobis D – odległość Mahalanobisa została wyznaczona na bazie 10 skal (np. skala przynależności, wartości samego siebie), z czego 6 było podskalami skali DOSPERT. Próg wartości odstających wyznaczono na poziomie 0,01 liczonego dla rozkładu χ^2. Średnia wielkość metryki wynosiła $M = 9{,}94$, $SD = 4{,}96$.

Korelacja Ogół-Osoba była wyznaczona na podstawie 57 pozycji przed przekodowaniem pozycji odwróconych. Próg ustalono na poziomie korelacji zerowej. Zdecydowana większość respondentów miała

współczynnik korelacji powyżej 50%. Średnia wielkość metryki wynosiła $M = 0{,}54$, $SD = 0{,}228$.

Metoda długich ciągów (*longstrings*) – wykorzystano 57 pozycji, przed przekodowaniem pozycji odwróconych, pochodzących z tych samych 5 skal co w przypadku wyznaczania odległości Mahalanobisa. Próg ustalono arbitralnie na podstawie analizy rozkładu wielkości na poziomie więcej niż 11 takich samych odpowiedzi. Niektóre wyniki o wysokich wartościach były spowodowane obecnością braku odpowiedzi. Średnia wielkość metryki wynosiła $M = 5{,}55$, $SD = 6{,}992$.

MAD (*mean verage differencea*) – była wyznaczona na podstawie trzech skal, w których występowały jednocześnie pozycje o standardowych i odwróconych wynikach (control, MeanExist, Belonging). Wartości bezwzględne uzyskane w każdej skali były uśrednione tak, aby utworzyć indeks. Wielkość progu nieuważności została wyznaczona na dwa odchylenia standardowe (2*SD*) od średniej. Średnia wielkość metryki wynosiła $M = 1{,}34$, $SD = 0{,}927$.

Stosując cztery wskazane powyżej metody wykrywania NNR, oznakowano relatywnie niewielkie odsetki osób uczestniczących w badaniu (zob. tabela 2). Nie licząc braków odpowiedzi, nieuważni i nietypowi respondenci stanowili maksymalnie 11,7%, co jest wynikiem bardzo dobrym.

Tabela 2

Liczebności NNR uzyskane w odtworzonym badaniu

Metoda	Longstrings	MAD (2*SD*)	Mahalanobis	Total Person
Longstrings	9			
MAD (2SD)	1	9		
Mahalanobis	0	0	3	
Total Person	1	0	0	4

Źródło: opracowanie własne.

Zastosowane metody były mało redundantne – co można zaobserwować, analizując wspólne części dla metod – najczęściej są to wielkości zerowe. Tylko w przypadku metody długich ciągów zaobserwowano powtarzanie się pojedynczych przypadków nieuważności z metodą MAD oraz metodą korelacji Ogół-Osoba.

Etap 3. Wyniki moderowanej mediacji przy zastosowaniu poszczególnych metod wykrywania NNR

Z perspektywy badawczej przedstawionej w Pfundmair i Lermer (2023) istotną rolą zajmuje moderowana mediacja. W opisywanym artykule była ona

istotna statystycznie. Udało się ją również uzyskać w badaniu odtworzonym (zob. tabela 1).

W trzecim etapie po kolei ograniczano liczbę respondentów związaną z poszczególnymi metodami wykrywania NNR. Wyniki nie ulegały zmianie w odniesieniu do efektów mediacji. Sytuacja była jednak inna w przypadku efektu moderowanej mediacji.

Tabela 3

Wyniki interakcji Dotkliwość*Włączenie –> Dominacja (Severity*Inlcusion –> Dominace) w zależności od wykorzystanej metody wykrywania NNR

Metoda	*b*	*SE*	*t*	*p*	*LL*	*UL*
Mahalanobis	–0,0996	0,0505	–1,9713	0,0504	–0,1994	0,0002
Total Person	0,0544	0,0544	–1,9491	0,053	–0,2135	0,0014
MAD	–0,1226	0,0563	–2,1790	0,0308	–0,2338	–,0115
Longstrings	–0,1006	0,0545	–1,8441	0,067	–0,2083	0,0071

Źródło: opracowanie własne na podstawie danych Pfundmair i Lermer (2023) z uwzględnieniem analiz NNR.

Zastosowanie metod wykrywania NNR spowodowało zanik interakcji pomiędzy zmiennymi *dotkliwość* i *dominacja*. Taka sytuacja zachodziła zarówno w przypadku użycia metody Mahalanobisa, korelacji Ogół-Osoba oraz metody długich ciągów (zob. tabela 2). We wszystkich tych przypadkach przedziały ufności (*LL*:*UL*) obliczone procedurą bootstrapingu zawierały zero (Hayes, 2018). Wyniki pozostały istotne statystycznie (na poziomie $\alpha = 0,05$) po zastosowaniu metody MAD. Generalnie więc zastosowanie metod wykrywających nietypowych respondentów i ich usunięcie prowadziło do zaniku efektu moderowanej mediacji. Natomiast usunięcie respondentów nieuważnych prowadziło do niejednoznacznych efektów. W kontekście przyszłych badań kluczowe mogłoby być określenie kryteriów determinujących wpływ tego efektu na pewne grupy osób, a brak wpływu na inne grupy badanych.

VII. DYSKUSJA I WNIOSKI

Ogólny wniosek, jaki można wyciągną z zaprezentowanego badania jest taki, że nie udało się odtworzyć wyników badania Pfundmair i Lermer (2023) w sytuacji, gdy wyeliminowano respondentów nieuważnych lub nietypowych. Należy zwrócić uwagę na fakt, że brak istotności statystycznej interakcji występował już po usunięciu tylko kilku respondentów, jak w przypadku metody Mahalanobisa (3 respondentów) czy korelacji Ogół-Osoba (4 badanych).

Warto rozważyć, nawet czysto hipotetycznie, szanse na replikację wyników takiego badania. Po pierwsze, można zauważyć, że gdyby w badaniu replika-

cyjnym zastosowano metody wykrywania NNR, z dużą dozą prawdopodobieństwa doprowadziłoby to do braku odtworzenia wyników oryginalnego badania. Za tą argumentacją stają wyniki uzyskane w prezentowanym badaniu, w którym nie uzyskano odtworzenia wyników na tych samych danych w przypadku użycia niektórych z metod wykrywania NNR. Replikacja wyników na innych danych byłaby zapewne jeszcze trudniejsza.

Po drugie, nawet jeżeli w badaniu replikacyjnym nie zastosowano by metod wykrywania NNR, to z niezerowym prawdopodobieństwem można założyć, że odtworzenie wyników byłoby trudne. Uzyskane w oryginalnym badaniu istotne statystycznie wyniki były częściowo związane z nielicznymi osobami nietypowymi. Trudno podejrzewać, że takie same osoby musiałby się pojawić w replikowanych badaniach.

Kwestią kluczową z poznawczego i etycznego punktu widzenia jest ta, czy efekt interakcji opisywany w artykule Pfundmair i Lermer (2023) zasługuje na jego przedstawienie. Odpowiedź na to pytanie jest oczywista, ale warunkowa. Zgodnie z sugestiami z literatury (Aguinis i in., 2018) wyniki, niezależnie od ich statystycznej istotności, należy przedstawić. Problem polega na tym, czy przedstawić je jako istotne (bez uwzględnienia metod wykrywania NNR) czy jako nieistotne statystycznie (z uwzględnieniem metod wykrywania NNR). Jak to często bywa, nie ma na nie jednoznacznej odpowiedzi. W jednym i drugim przypadku narażonym się jest na możliwość popełnienia albo błędu I typu, albo błędu II typu. Jedynym rozsądnym rozwiązaniem w tym przypadku wydaje się przedstawianie pełnych wyników, ze wskazaniem wrażliwości uzyskanych efektów na występowanie przypadków NNR. Taka analiza wrażliwości efektów w zależności od stosowanych metod NNR oznaczałaby przedstawienie nie jednego, ale wielu wyników. Nawet jeżeli generalnie ludzie nie lubią niejasności (Fox i Tversky, 1995), to rozwiązanie takie pozwalałoby innym zespołom badaczy na zwiększenie szans na udane replikowanie badań. Działoby się to dzięki zwiększeniu komparatywności prób badawczych, czyli wykorzystaniu takiego samego zbioru metod NNR albo wykorzystaniu określonej zmiennej (np. poziomu uwagi) jako zmiennej moderującej.

Niebagatelną rolę odgrywa też udostępnianie danych. Daje ono możliwość, nawet *ex post*, sprawdzenia poziomu typowości lub uważności respondentów we wcześniejszych badaniach. Gdyby w replikowanym badaniu nie uzyskano podobnych rezultatów, to zawsze dzięki dostępowi do oryginalnych danych będzie istniała możliwość znalezienia źródła niepowodzenia. Może nim być ingerencja w dane związana z usunięciem NNR.

Klamrą spinającą wszystkie wymienione elementy jest kwestia pełnego raportowania procedur i wyników badania. Dotyczy to też metod, progów i sposobów postępowania z nietypowymi i nieuważnymi respondentami.

Przedstawione powyżej wnioski można zaprezentować w formie zaleceń, co należałoby robić, aby zwiększyć szanse na replikację badania:

1. Stosować metody wykrywania NNR.

2. Rejestrować przed badaniami zakres użytych metod wykrywania NNR, czyli określać typy wykorzystywanych metod, progi i procedury postępowania z wykrytymi przypadkami NNR.

3. Zapewnić dostęp do danych innym zespołom badaczy.

4. Określić wrażliwość uzyskanych efektów w zależności od eliminacji różnych grup respondentów NNR.

5. Raportować w ostatecznej wersji publikacji o postępowaniu w zakresie NNR.

VIII. PODSUMOWANIE

W artykule postawiono za cel przedstawienie problemu powstającego na styku replikacji oraz metod związanych z eliminacją nieuważnych i nietypowych respondentów. Przedstawione wyniki dotyczące odtworzenia badania z wykorzystaniem metod eliminujących nieuważnych lub nietypowych respondentów dostarczają argumentów na niemożliwość odtworzenia wyników oryginalnego badania. Z uwagi na fakt, że w odtworzonym badaniu eliminowane były bardzo niewielkie liczebności respondentów, zanik oryginalnych efektów jest szczególnie istotny. Oznacza on, że nawet bardzo małe zmiany w strukturze danych mogą prowadzić do braku replikacji. W ostatecznym więc rachunku metody wykrywania nietypowych i nieuważnych respondentów mogą potęgować kryzys replikacyjności.

Z metod wykrywania nietypowych i nieuważnych respondentów nie można jednak zrezygnować (Ward i Meade, 2023). Ich stosowanie daje szansę na zwiększenie komparatywności prób badawczych. Jest to zgodne z przyjętym w literaturze kanonem postępowania. Stosowanie metod wykrywania NNR nie jest jednak lekiem na całe zło. Daje bowiem możliwość manipulowania zestawem danych. Zabezpieczyć się przed tym można wcześniejszą rejestracją programu badawczego. Co prawda wcześniejsza rejestracja metod, progów i procedur wykrywania i usuwania NNR zdejmuje z badaczy odium posługiwania się wątpliwymi praktykami badawczymi, ale jednak nie rozwiązuje problemu jako całości. Po pierwsze, nie zawsze możliwe jest wcześniejsze określenie progów dla wybranych metod. Te wynikają często z charakterystyki uzyskanych rezultatów i choćby liczby nieuważnych i nietypowych respondentów. Po drugie, tylko częściowo redukuje problemy z replikacją badania. Na pewno tak się stanie w przypadku braku raportowania o użytych metodach wykrywania NNR. Nawet jednak w przypadku raportowania replikacja będzie bardziej prawdopodobna, niemniej może dotyczyć tylko specyficznej sytuacji badawczej – przykładowo, próby, z której usunie się na przykład dużą część nietypowych respondentów. Problem polega na tym, czy taki „sztukowany" efekt będzie bliżej prawdy, czy tylko będzie sprytną sztuczką replikacji nieprawdy.

W kontekście problemów z replikacją bardzo cenne wydaje się udostępnianie danych innym zespołom badaczy. W analizowanym przypadku dane te pozwoliły na napisanie artykułu, a w innych przypadkach dałyby możliwość przetestowania efektów w sytuacji, gdy w replikacji nie zanotowano istotnych statystycznie efektów. Wtedy możliwość sprawdzenia, czy wynika to z obecności NNR, daje prostą i szybką odpowiedź na to pytanie.

Wspomniane powyżej działania mieszczą się w ramach zgłaszanych do tej pory propozycji dotyczących poprawienia jakości badań naukowych (Open Science Collaboration, 2015). Warto wspomnieć o innej możliwości, która w omawianym kontekście mogłaby mieć istotne znaczenie. Chodzi mianowicie o prezentowanie wyników badania (lub choćby skrótowej informacji) z uwzględnieniem różnych typów metod wykrywania NNR. W takim przypadku pojawiałaby się informacja o tym, że wykryte efekty występują lub zanikają w zależności od struktury próby. Informowałoby to innych badaczy o stabilności uzyskiwanych efektów oraz wskazywałoby na potencjalne istnienie istotnych czynników moderujących (jak poziom zaangażowania czy potrzeby poznania). Chroniłoby również autorów przed zarzutem p-hackingu, szczególnie w sytuacji, gdy niektóre z analizowanych wariantów wyników prowadziły do nieistotnych statystycznie rezultatów.

Bibliografia

Agnoli, F., Wicherts, J. M., Veldkamp, C. L., Albiero, P., i Cubelli, R. (2017). Questionable research practices among Italian research psychologists. *PloS ONE*, *12*(3), e0172792. https://doi.org/10.1371/journal.pone.0172792

Aguinis, H., Ramani, R. S., i Alabduljader, N. (2018). What you see is what you get? Enhancing methodological transparency in management research. *Academy of Management Annals*, *12*(1), 83–110. https://doi.org/10.5465/annals.2016.0011

Amrhein, V., Trafimow, D., i Greenland, S. (2019). Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. *The American Statistician*, *73*(sup1), 262–270. https://doi.org/10.1080/00031305.2018.1543137

Andrade, C. (2021). HARKing, cherry-picking, p-hacking, fishing expeditions, and data dredging and mining as questionable research practices. *The Journal of Clinical Psychiatry*, *82*(1), e1–e3. https://doi.org/10.4088/JCP.20f13804

Baillie, M., Le Cessie, S., Schmidt, C. O., Lusa, L., Huebner, M., for the Topic Group "Initial Data Analysis" of the STRATOS Initiative. (2022). Ten simple rules for initial data analysis. *PLoS Computational Biology*, *18*(2), e1009819. https://doi.org/10.1371/journal.pcbi.1009819

Bakker, B. N., Jaidka, K., Dörr, T., Fasching, N., i Lelkes, Y. (2021). Questionable and open research practices: Attitudes and perceptions among quantitative communication researchers. *Journal of Communication*, *71*(5), 715–738. https://doi.org/10.1093/joc/jqab031

Banks, G. C., O'Boyle, E. H., Pollack, J. M., White, C. D., Batchelor, J. H., Whelpley, C. E., Abston, K. A., Bennett, A. A., i Adkins, C. L. (2016). Questions about questionable research practices in the field of management: A guest commentary. *Journal of Management*, *42*(1), 5–20. https://doi.org/10.1177/0149206315619011

Benjamini, Y. (2020). Selective inference: The silent killer of replicability. *Harvard Data Science Review*, *2*(4). https://doi.org/10.1162/99608f92.fc62b261

Białek, A., i Wolski, P. (2023). Dwa głosy o kryzysie wiarygodności w psychologii. *Przegląd Psychologiczny*, *66*(1), 9–26. https://doi.org/10.31648/przegldpsychologiczny.9455

Bridges, A. J. (2022). Hypothesizing after results are known: HARKing. W W. O'Donohue, A. Masuda i S. Lilienfeld (red.), *Avoiding questionable research practices in applied psychology* (s. 175–190). Springer International Publishing.

Brzeziński, J. M. (2023). Czy kryzys wiarygodności w psychologii? *Przegląd Psychologiczny*, *66*(1), 27–47. https://doi.org/10.31648/przegldpsychologiczny.9456

Bybee, S., Cloyes, K., Baucom, B., Supiano, K., Mooney, K., i Ellington, L. (2022). Bots and nots: Safeguarding online survey research with underrepresented and diverse populations. *Psychology Sexuality*, *13*(4), 901–911. https://doi.org/10.1080/19419899.2021.1936617

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., … Wu, H. (2018). Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z

Cockburn, A., Dragicevic, P., Besançon, L., i Gutwin, C. (2020). Threats of a replication crisis in empirical computer science. *Communications of the ACM*, *63*(8), 70–79. https://doi.org/10.1145/3360311

Coiera, E., Ammenwerth, E., Georgiou, A., i Magrabi, F. (2018). Does health informatics have a replication crisis? *Journal of the American Medical Informatics Association*, *25*(8), 963–968. https://doi.org/10.1093/jamia/ocy028

Cornell, D., Klein, J., Konold, T., i Huang, F. (2012). Effects of validity screening items on adolescent survey data. *Psychological Assessment*, *24*(1), 21–35. https://doi.org/10.1037/a0024824

Credé, M. (2010). Random responding as a threat to the validity of effect size estimates in correlational research. *Educational and Psychological Measurement*, *70*(4), 596–612. https://doi.org/10.1177/0013164410366686

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4–19. https://doi.org/10.1016/j.jesp.2015.07.006

DeSimone, J. A., i Harms, P. D. (2018). Dirty data: The effects of screening respondents who provide low-quality data in survey research. *Journal of Business and Psychology*, *33*(5), 559–577. https://doi.org/10.1007/s10869-017-9514-9

Dunn, A. M., Heggestad, E. D., Shanock, L. R., i Theilgard, N. (2018). Intra-individual response variability as an indicator of insufficient effort responding: Comparison to other indicators and relationships with individual differences. *Journal of Business and Psychology*, *33*(1), 105–121. https://doi.org/10.1007/s10869-016-9479-0

Duvendack, M., Palmer-Jones, R., i Reed, W. R. (2017). What is meant by "replication" and why does it encounter resistance in economics? *American Economic Review*, *107*(5), 46–51. https://doi.org/10.1257/aer.p20171031

Ellis, R. J. (2022). Questionable research practices, low statistical power, and other obstacles to replicability: Why preclinical neuroscience research would benefit from registered reports. *Eneuro*, *9*(4), ENEURO.0017-22.2022. https://doi.org/10.1523/ENEURO.0017-22.2022

Fox, C. R., i Tversky, A. (1995). Ambiguity aversion and comparative ignorance. *The Quarterly Journal of Economics*, *110*(3), 585–603. https://doi.org/10.2307/2946693

Franco, A., Malhotra, N., i Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505. https://www.science.org/doi/epdf/10.1126/science.1255484

Franco, A., Malhotra, N., i Simonovits, G. (2016). Underreporting in psychology experiments: Evidence from a study registry. *Social Psychological and Personality Science*, *7*(1), 8–12. https://doi.org/10.1177/1948550615598377

Freese, J., i Peterson, D. (2017). Replication in social science. *Annual Review of Sociology*, *43*(1), 147–165. https://doi.org/10.1146/annurev-soc-060116-053450

Hayes, A. F. (2018). *Introduction to mediation: A regression-based approach*. Guilford Press.

He, J., Van de Vijver, F. J., Espinosa, A. D., i Mui, P. H. (2014). Toward a unification of acquiescent, extreme, and midpoint response styles: A multilevel study. *International Journal of Cross Cultural Management*, *14*(3), 306–322. https://doi.org/10.1177/1470595814541424

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., i DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*(1), 99–114. https://doi.org/10.1007/s10869-011-9231-8

Hudson, R. (2023). Explicating exact versus conceptual replication. *Erkenntnis*, *88*(6), 2493–2514. https://doi.org/10.1007/s10670-021-00464-z

Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*, 640–648. https://doi.org/10.1097/EDE.0b013e31818131e7

Kam, C. C. S., Meyer, J. P. (2015). How careless responding and acquiescence response bias can influence construct dimensionality: The case of job satisfaction. *Organizational Research Methods*, *18*(3), 512–541. https://doi.org/10.1177/1094428115571894

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236. https://doi.org/10.1002/acp.2350050305

Lakens, D. (2015). On the challenges of drawing conclusions from *p*-values just below 0.05. *PeerJ*, *3*, e1142. https://peerj.com/articles/1142/

Lakens, D., i Evers, E. R. K. (2014). Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science*, *9*(3), 278–292. https://doi.org/10.1177/1745691614528520

Leichtmann, B., Nitsch, V., i Mara, M. (2022). Crisis ahead? Why human-robot interaction user studies may have replicability problems and directions for improvement. *Frontiers in Robotics and AI*, *9*, 838116. https://doi.org/10.3389/frobt.2022.838116

Leys, C., Delacre, M., Mora, Y. L., Lakens, D., i Ley, C. (2019). How to classify, detect, and manage univariate and multivariate outliers, with emphasis on pre-registration. *International Review of Social Psychology*, *32*(1, Article 5. https://doi.org/10.5334/irsp.289

Lynch Jr, J. G., Bradlow, E. T., Huber, J. C., i Lehmann, D. R. (2015). Reflections on the replication corner: In praise of conceptual replications. *International Journal of Research in Marketing*, *32*(4), 333–342. https://doi.org/10.1016/j.ijresmar.2015.09.006

Maniaci, M. R., i Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, *48*, 61–83. https://doi.org/10.1016/j.jrp.2013.09.008

Maxwell, S. E., Lau, M. Y., i Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*(6), 487–498. https://doi.org/10.1037/a0039400

Meade, A. W., i Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. https://doi.org/10.1037/a0028085

Muradchanian, J., Hoekstra, R., Kiers, H., i van Ravenzwaaij, D. (2023). The role of results in deciding to publish: A direct comparison across authors, reviewers, and editors based on an online survey. *PloS ONE, 18*(10), e0292279. https://doi.org/10.1371/journal.pone.0292279

Muthukrishna, M., Bell, A. V., Henrich, J., Curtin, C. M., Gedranovich, A., McInerney, J., i Thue, B. (2020). Beyond Western, educated, industrial, rich, and democratic (WEIRD) psychology: Measuring and mapping scales of cultural and psychological distance. *Psychological Science*, *31*(6), 678–701. https://doi.org/10.1177/0956797620916782

Niessen, A. S. M., Meijer, R. R., i Tendeiro, J. N. (2016). Detecting careless respondents in web-based questionnaires: Which method to use? *Journal of Research in Personality*, *63*, 1–11. https://doi.org/10.1016/j.jrp.2016.04.010

Nosek, B. A., i Errington, T. M. (2020). What is replication? *PLoS Biology*, *18*(3), e3000691. https://doi.org/10.1371/journal.pbio.3000691

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., i Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, *73*(1), 719–748. https://doi.org/10.1146/annurev-psych-020821-114157

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Oppenheimer, D. M., Meyvis, T., i Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867–872. https://doi.org/10.1016/j.jesp.2009.03.009

Pagell, M. (2021). Replication without repeating ourselves: Addressing the replication crisis in operations and supply chain management research. *Journal of Operations Management*, *67*(1), 105–115. https://doi.org/10.1002/joom.1120

Peels, R., i Bouter, L. (2018). The possibility and desirability of replication in the humanities. *Palgrave Communications*, *4*(1), 95. https://doi.org/10.1057/s41599-018-0149-x

Pfundmair, M, i Lermer, E. (2023) Examining the link between social exclusion and social-risk taking: A correlational and experimental investigation [version 1; peer review: 1 approved

with reservations], *Routledge Open Research*, 2023, 2:4 Last updated: 28 Mar 2023. https://osf.io/umrht

Ravn, T., i Sørensen, M. P. (2021). Exploring the gray area: Similarities and differences in questionable research practices (QRPs) across main areas of research. *Science and Engineering Ethics*, *27*(4), 40. https://doi.org/10.1007/s11948-021-00310-z

Richard, F. D., Bond, C. F., i Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7*(4), 331–363. https://doi.org/10.1037/1089-2680.7.4.331

Saad, D. (2021). Nowe narzędzia i techniki zwiększające trafność badań internetowych. *Com.Press*, *4*(1), 106–121. https://doi.org/10.51480/compress.2021.4-1.248

Scheel, A. M., Schijen, M. R. M. J., i Lakens, D. (2021). An excess of positive results: Comparing the standard psychology literature with registered reports. *Advances in Methods and Practices in Psychological Science*, *4*(2), 1–12. https://doi.org/10.1177/25152459211007467

Stodden, V. (2015). Reproducing Statistical Results. *Annual Review of Statistics and Its Application*, *2*(1), 1–19. https://doi.org/10.1146/annurev-statistics-010814-020127

Świątkowski, W., i Dompnier, B. (2017). Replicability crisis in social psychology: Looking at the past to find new pathways for the future. *International Review of Social Psychology*, *30*(1), 111–124. https://doi.org/10.5334/irsp.66

Tincani, M., i Travers, J. (2019). Replication research, publication bias, and applied behavior analysis. *Perspectives on Behavior Science*, *42*(1), 59–75. https://doi.org/10.1007/s40614-019-00191-5

Vecchio, R., Caso, G., Cembalo, L., i Borrello, M. (2020). Is respondents' inattention in online surveys a major issue for research? *Economia Agro-Alimentare*, *1*, 1–18. https://doi.org/10.3280/ecag1-2020oa10069

Vowels, M. J. (2021). Misspecification and unreliable interpretations in psychology and social science. *Psychological Methods*, 28, 507–526. https://doi.org/10.1037/met0000429

Wanat, T. (2010). Niedoskonałości w formulowaniu hipotez badawczych w pracach doktorskich. *Zagadnienia Naukoznawstwa*, *46*(1), 27–41.

Ward, M., i Meade, A. W. (2023). Dealing with careless responding in survey data: Prevention, identification, and recommended best practices. *Annual Review of Psychology*, *74*, 577–596. https://doi.org/10.1146/annurev-psych-040422-045007

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., Van Aert, R. C. M., i Van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*, 1832. https://doi.org/10.3389/fpsyg.2016.01832

Yang, S., Wang, L., Ding, P. (2019). Causal inference with confounders missing not at random. *Biometrika*, *106*(4), 875–888. https://doi.org/10.1093/biomet/asz048